THE TEXT RETRIEVAL CONFERENCES (TRECS)

Ellen M. Voorhees, Donna Harman
National Institute of Standards and Technology
Gaithersburg, MD 20899

1 INTRODUCTION

Phase III of the TIPSTER project included three workshops for evaluating document detection (information retrieval) projects: the fifth, sixth and seventh Text REtrieval Conferences (TRECs). This work was co-sponsored by the National Institute of Standards and Technology (NIST), and included evaluation not only of the TIPSTER contractors, but also of many information retrieval groups outside of the TIPSTER project. The conferences were run as workshops that provided a forum for participating groups to discuss their system results on the retrieval tasks done using the TIPSTER/TREC collection. As with the first four TRECs, the goals of these workshops were:

- to encourage research in text retrieval based on large test collections;
- to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems;
- to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems; and
- to serve as a showcase for state-of-the-art retrieval systems for DARPA and its clients.

For each TREC, NIST provides a test set of documents and questions. Participants run their retrieval systems on the data, and return to NIST a list of the retrieved top-ranked documents. NIST pools the individual results, judges the retrieved documents for correctness, and evaluates the results. The TREC cycle ends with a workshop that is a forum for participants to share their experiences. The most recent workshop in the series, TREC-7, was held at NIST in November 1998.

The number of participating systems has grown from 25 in TREC-1 to 38 in TREC-5 (Table 1), 51 in TREC-6 (Table 1), and 56 in TREC-7 (Table 1). The groups include representatives from 16 different countries and 32 companies.

TREC provides a common test set to focus research on a particular retrieval task, yet actively encourages participants to do their own experiments within the umbrella task. The individual experiments broaden the scope of the research that is done within TREC and make TREC more attractive to individual participants. This marshaling of research efforts has succeeded in improving the state of the art in retrieval technology, both in the level of basic performance (see Figure 1) and in the ability of these systems to function well in diverse environments, such as retrieval in a filtering operation or retrieval against multiple languages.

Each of the TREC conferences has centered around two main tasks: the routing task (not run in TREC-7) and the ad hoc task (these tasks are described in more detail in Section 2.3). In addition, starting in TREC-4 a set of "tracks" or tasks that focus on particular subproblems of text retrieval was introduced. These tracks include tasks that concentrate on a specific part of the retrieval process (such as the interactive track which focuses on user-related issues), or tasks that tackle research in related areas, such as the retrieval of spoken "documents" from news broadcasts.

The graph in Figure 1 shows that retrieval effectiveness has approximately doubled since the beginning of TREC. This means, for example, that retrieval engines that could retrieve three good documents within the top ten documents retrieved in 1992 are now likely to retrieve six good documents in the top ten documents retrieved for the same search. The figure plots retrieval effectiveness for one well-known retrieval engine, the SMART system of Cornell University. The SMART system has consistently been one of the more effective systems in TREC, but other systems are

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **OCT 1998**	2. REPORT TYPE	3. DATES COVERED **00-00-1998 to 00-00-1998**

4. TITLE AND SUBTITLE **The Text Retrieval Conferences (TRECs)**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **National Institute of Standards and Technology,Gaithersburg,MD,20899**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998. Sponsored by the Defense Advanced Research Projects Agency.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**33**	

Apple Computer
Australian National University
CLARITECH Corporation
City University
Computer Technology Institute
Cornell University
Dublin City University
FS Consulting
GE/NYU/Rutgers/Lockheed Martin
GSI-Erli
George Mason University
IBM Corporation
IBM T.J. Watson Research Center
Information Technology Institute, Singapore
Institut de Recherche en Informatique de Toulouse
Intext Systems
Lexis-Nexis
MDS at RMIT
MITRE
Monash University
New Mexico State University (two groups)
Open Text Corporation
Queens College, CUNY
Rank Xerox Research Center
Rutgers University (two groups)
Swiss Federal Institute of Technology (ETH)
Universite de Neuchatel
University of California, Berkeley
University of California, San Diego
University of Glasgow
University of Illinois at Urbana-Champaign
University of Kansas
University of Maryland
University of Massachusetts, Amherst
University of North Carolina
University of Waterloo

Table 1: TREC-5 participants

Apple Computer
AT&T Labs Research
Australian National University
CEA (France)
Carnegie Mellon University
Center for Information Research, Russia
City University, London
CLARITECH Corporation
Cornell U./SaBIR Research, Inc
CSIRO (Australia)
Daimler Benz Research Center Ulm
Dublin City University
Duke U./U. of Colorado/Bellcore
FS Consulting, Inc.
GE Corp./Rutgers U.
George Mason U./NCR Corp.
Harris Corp.
IBM T.J. Watson Research (2 groups)
ITI (Singapore)
MSI/IRIT/U. Toulouse (France)
ISS (Singapore)
APL, Johns Hopkins University
Lexis-Nexis
MDS at RMIT, Australia
MIT/IBM Almaden Research Center
NEC Corporation
New Mexico State U. (2 groups)
NSA (Speech Research Branch)
Open Text Corporation
Oregon Health Sciences U.
Queens College, CUNY
Rutgers University (2 groups)
Siemens AG
SRI International
Swiss Federal Inst. of Tech.(ETH)
TwentyOne (TNO/U-Tente/DFKI/Xerox/U-Tuebingen)
U. of California, Berkeley
U. of California, San Diego
U. of Glasgow
U. of Maryland, College Park
U. of Massachusetts, Amherst
U. of Montreal
U. of North Carolina (2 groups)
U. of Sheffield/U. of Cambridge
U. of Waterloo
Verity, Inc.
Xerox Research Centre Europe

Table 2: TREC-6 participants

ACSys Cooperative Research Centre
AT&T Labs Research
Avignon CS Laboratory/Bertin
BBN Technologies
Canadian Imperial Bank of Commerce
Carnegie Mellon University
Commissariat à l'Energie Atomique
CLARITECH Corporation
Cornell University/SabIR Research, Inc.
Defense Evaluation and Research Agency
Eurospider
Fondazione Ugo Bordoni
FS Consulting, Inc.
Fujitsu Laboratories, Ltd.
GE/Rutgers/SICS/Helsinki
Harris Information Systems Division
IBM — Almaden Research Center
IBM T.J. Watson Research Center (2 groups)
Illinois Institute of Technology
Imperial College of Science, Technology and Medicine
Institut de Recherche en Informatique de Toulouse
The Johns Hopkins University — APL
Kasetsart University
KDD R&D Laboratories
Keio University
Lexis-Nexis
Los Alamos National Laboratory
Management Information Technologies, Inc.
Massachusetts Institute of Technology
National Tsing Hua University
NEC Corp. and Tokyo Institute of Technology
New Mexico State University
NTT DATA Corporation
Okapi Group (City U./U. of Sheffield/Micr osoft)
Oregon Health Sciences University
Queens College, CUNY
RMIT/Univ. of Melbourne/CSIRO
Rutgers University (2 groups)
Seoul National University
Swiss Federal Institute of Technology (ETH)
TextWise, Inc.
TNO-TPD TU-Delft
TwentyOne
Universite de Montreal
University of California, Berkeley
University of Cambridge
University of Iowa
University of Maryland
University of Massachusetts, Amherst
University of North Carolina, Chapel Hill
Univ. of Sheffield/Cambridge/SoftSound
University of Toronto
University of Waterloo
U.S. Department of Defense

Table 3: TREC-7 participants

comparable with it, so the graph is representative of the increase in effectiveness for the field as a whole.

Researchers at Cornell ran the version of SMART used in each of the seven TREC conferences against each of the seven ad hoc test sets (Buckley, Mitra, Walz, & Cardie, 1999). Each line in the graph connects the mean average precision scores produced by each version of the system for a single test. For each test, the TREC-7 system has a markedly higher mean average precision than the TREC-1 system. The recent decline in the absolute scores reflects the evolution towards more realistic, and difficult, test questions, and also possibly a dilution of effort because of the many tracks being run in TRECs 5, 6, and 7.

The seven TREC conferences represent hundreds of retrieval experiments. The Proceedings of each conference captures the details of the individual experiments, and the Overview paper in each Proceedings summarizes the main findings of each conference. A special issue on TREC-6 will be published in *Information Processing and Management* (Voorhees, in press), which includes an Overview of TREC-6 (Voorhees & Harman, in press) as well as an analysis of the TREC effort by Sparck Jones (in press).

2 THE TASKS

Each of the TREC conferences has centered around two main tasks, the routing task and the ad hoc task. In addition, starting in TREC-4 a set of "tracks," tasks that focus on particular subproblems of text retrieval, was introduced. This section describes the goals of the two main tasks. Details regarding the tracks are given in Section 6.

2.1 The Routing Task

The routing task in the TREC workshops investigates the performance of systems that use standing queries to search new streams of documents. These searches are similar to those required by news clipping services and library profiling systems. A true routing

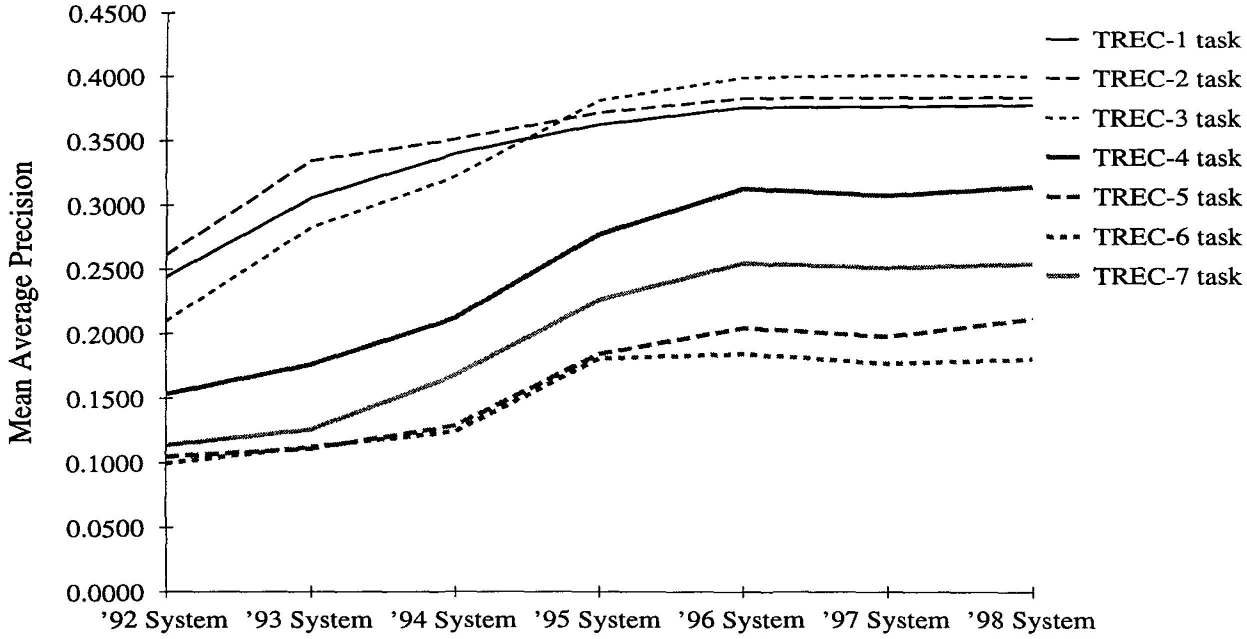

Figure 1: Retrieval effectiveness improvement for Cornell's SMART system, TREC-1 - TREC-7.

environment is simulated in TREC by using questions (called *topics* in TREC) for which the right set of documents to be retrieved is known for one document set, and then testing the systems' performance with those questions on a completely new document set.

The training for the routing task is shown in the left-hand column of Figure 2. Participants are given a set of topics and a document set that includes known relevant documents for those topics. The topics consist of natural language text describing a user's information need (see sec. 3.2 for details). The topics are used to create a set of queries (the actual input to the retrieval system) that are then used against the training documents. This is represented by Q1 in the diagram. Many Q1 query sets might be built to help adjust the retrieval system to the task, to create better weighting algorithms, and to otherwise prepare the system for testing. The result of the training is query set Q2, routing queries derived from the routing topics and run against the test documents.

The testing phase of the routing task is shown in the middle column of Figure 2. The output of running Q2 against the test documents is the official test result for the routing task.

2.2 The Ad Hoc Task

The ad hoc task investigates the performance of systems that search a static set of documents using new topics. This task is similar to how a researcher might use a library—the collection is known but the questions likely to be asked are not known. The right-hand column of Figure 2 depicts how the ad hoc task is accomplished in TREC. Participants are given a document collection consisting of approximately 2 gigabytes of text and 50 new topics. The set of relevant documents for these topics in the document set is not known at the time the participants receive the topics. Participants produce a new query set, Q3, from the ad hoc topics and run those queries against the ad hoc documents. The output from this run is the official test result for the ad hoc task.

2.3 Task Guidelines

In addition to the task definitions, TREC participants are given a set of guidelines outlining acceptable methods of indexing, knowledge base construction, and generating queries from the supplied topics. In general, the guidelines are constructed to reflect an actual operational environment and to allow fair comparisons among the diverse query construction approaches. The allowable query construction methods in TRECs 5, 6, and 7 were divided into *au-*

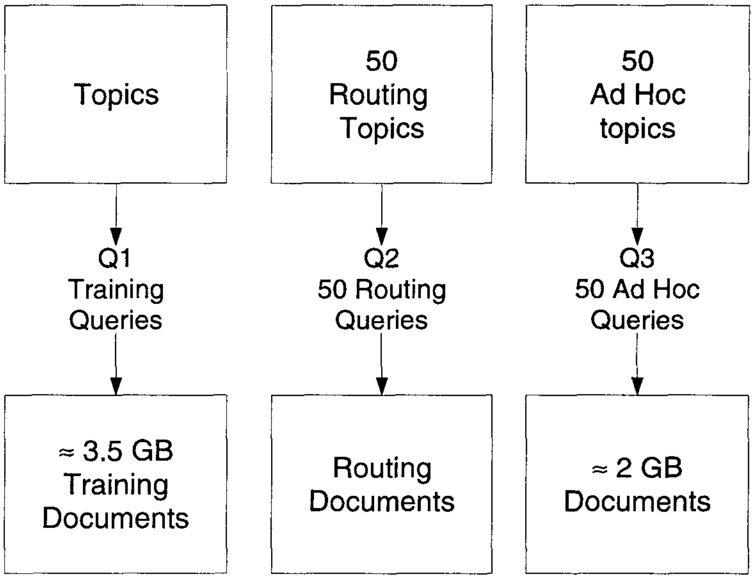

Figure 2: TREC main tasks.

tomatic methods, in which queries are derived completely automatically from the topic statements, and *manual* methods, which includes queries generated by all other methods. This definition of manual query construction methods permitted users to look at individual documents retrieved by the ad hoc queries and then reformulate the queries based on the documents retrieved.

3 THE TEST COLLECTIONS

Like most traditional retrieval test collections, there are three distinct parts to the collections used in TREC: the documents, the questions or topics, and the relevance judgments or "right answers." This section describes each of these pieces for the collections used in the main tasks in TRECs 5, 6, and 7. Many of the tracks have used the same data or used data constructed in a similar method but in a different environment, such as in multiple languages or using different guidelines (such as high precision searching).

3.1 Documents

TREC documents are distributed on CD-ROM's with approximately 1 GB of text on each, compressed to fit. Table 3.1 shows the statistics for all the English document collections used in TREC. TREC-5 used disks 2 and 4 for the ad hoc testing, while TRECs 6 and 7 used disks 4 and 5 for ad hoc testing. The FBIS on disk 5 (FBIS-1) was used for testing in the TREC-5 routing task and for training in the TREC-6 routing task, with new FBIS (FBIS-2) being used for testing in TREC-6. There was no routing task in TREC-7.

Documents are tagged using SGML to allow easy parsing (see Fig. 3). The documents in the different datasets have been tagged with identical major structures, but they have different minor structures. The philosophy in the formatting at NIST is to leave the data as close to the original as possible. No attempt is made to correct spelling errors, sentence fragments, strange formatting around tables, or similar faults.

3.2 Topics

In designing the TREC task, there was a conscious decision made to provide "user need" statements rather than more traditional queries. Two major issues were involved in this decision. First, there was a desire to allow a wide range of query construction methods by keeping the topic (the need statement) distinct from the query (the actual text submitted to the system). The second issue was the ability to increase the amount of information available about each topic, in particular to include with each topic a clear statement of what criteria make a document relevant.

The topics used in TREC-1 and TREC-2 (topics 1–150) were very detailed, containing multiple fields and lists of concepts related to the subject of the topics. The ad hoc topics used in TREC-3 (151–200)

	Size	#	Median #	Mean #
	(megabytes)	Docs	Words/Doc	Words/Doc
Disk 1				
Wall Street Journal, 1987–1989	267	98,732	245	434.0
Associated Press newswire, 1989	254	84,678	446	473.9
Computer Selects articles, Ziff-Davis	242	75,180	200	473.0
Federal Register, 1989	260	25,960	391	1315.9
abstracts of U.S. DOE publications	184	226,087	111	120.4
Disk 2				
Wall Street Journal, 1990–1992 (WSJ)	242	74,520	301	508.4
Associated Press newswire (1988) (AP)	237	79,919	438	468.7
Computer Selects articles, Ziff-Davis (ZIFF)	175	56,920	182	451.9
Federal Register (1988) (FR88)	209	19,860	396	1378.1
Disk 3				
San Jose Mercury News, 1991	287	90,257	379	453.0
Associated Press newswire, 1990	237	78,321	451	478.4
Computer Selects articles, Ziff-Davis	345	161,021	122	295.4
U.S. patents, 1993	243	6,711	4445	5391.0
Disk 4				
the *Financial Times*, 1991–1994 (FT)	564	210,158	316	412.7
Federal Register, 1994 (FR94)	395	55,630	588	644.7
Congressional Record, 1993 (CR)	235	27,922	288	1373.5
Disk 5				
Foreign Broadcast Information Service (FBIS-1)	470	130,471	322	543.6
the *LA Times*	475	131,896	351	526.5
TREC-6 Routing Test Data				
Foreign Broadcast Information Service (FBIS-2)	490	120,653	348	581.3

Table 4: Document collection statistics. Words are strings of alphanumeric characters. No stop words were removed and no stemming was performed.

were much shorter and did not contain the complex structure of the earlier topics. Nonetheless, participants in TREC-3 felt that the topics were still too long compared with what users normally submit to operational retrieval systems. Therefore the TREC-4 topics (201–250) were made even shorter: a single field consisting of a one sentence description of the information need. Figure 4 gives a sample topic from each of these sets.

One of the conclusions reached in TREC-4 was that the much shorter topics caused both manual and automatic systems trouble, and that there were issues associated with using short topics in TREC that needed further investigation (Harman, 1996). Accordingly, the TREC-5 ad hoc topics re-introduced the title and narrative fields, making the topics similar in format to the TREC-3 topics. TREC-6 and TREC-7 topics used this same format, as shown in Figure 5. While having the same format as the TREC-3 topics, on average the later topics are shorter (contain fewer words) than the TREC-3 topics. Table 3.2 shows the lengths of the various sections in the TREC topics as they have evolved over the 7 TRECs.

Since TREC-3, the ad hoc topics have been created by the same person (or *assessor*) who performed the relevance assessments for that topic. Each assessor comes to NIST with ideas for topics based on his or her own interests, and searches the ad hoc collection (looking at approximately 100 documents per topic) to estimate the likely number of relevant documents per candidate topic. NIST personnel select the final 50 topics from among these candidates, based on having both a reasonable range of estimated number of relevant documents across topics and on balancing the load across assessors.

3.3 Relevance Assessments

Relevance judgments are of critical importance to a test collection. For each topic it is necessary to compile a list of relevant documents—as comprehensive

```
<DOC>
<DOCNO>FT911-3</DOCNO>
<PROFILE>AN-BE0A7AAIFT</PROFILE>
<DATE>910514
</DATE>
<HEADLINE>
FT 14 MAY 91 / International Company News:  Contigas plans DM900m east German
project
</HEADLINE>
<BYLINE>
By DAVID GOODHART
</BYLINE>
<DATELINE>
BONN
</DATELINE>
<TEXT>
CONTIGAS, the German gas group 81 per cent owned by the utility Bayernwerk, said
yesterday that it intends to invest DM900m (Dollars 522m) in the next four years
to build a new gas distribution system in the east German state of Thuringia.  ...
</TEXT>
</DOC>
```

Figure 3: A document extract from the *Financial Times*.

a list as possible. All TRECs have used the pooling method (Sparck Jones & van Rijsbergen, 1975) to assemble the relevance assessments. In this method a pool of possible relevant documents is created by taking a sample of documents selected by the various participating systems. This pool is then shown to the human assessors. The particular sampling method used in TREC is to take the top 100 documents retrieved in each submitted run for a given topic and merge them into the pool for assessment. This is a valid sampling technique since all the systems used ranked retrieval methods, with those documents most likely to be relevant returned first. On average, an assessor judges approximately 1500 documents per topic.

Given the vital role relevance judgments play in a test collection, it is important to assess the quality of the judgments created in TREC. In particular, both the *completeness* and the *consistency* of the relevance judgments are of interest. Completeness measures the degree to which all the relevant documents for a topic have been found; consistency measures the degree to which the assessor has marked all the "truly" relevant documents relevant and the "truly" irrelevant documents irrelevant.

The completeness of the TREC relevance judgments has been investigated both at NIST (Harman, 1996) and independently at the Royal Melbourne Institute of Technology (RMIT) (Zobel, 1998). Both studies found that the completeness for most topics is adequate, though topics with many relevant documents are likely to have yet more relevant documents that have not been found through pooling. For this reason, NIST has deliberately chosen more tightly focused topics in recent TRECs. Both studies also found that any lack of completeness did not bias the results of particular systems. Indeed, the RMIT study showed that systems that did not contribute documents to the pool can still be evaluated fairly with the resulting judgments.

The consistency of the TREC judgments was investigated at NIST by obtaining multiple independent assessments for a set of topics and evaluating systems using each of the different judgment sets (Voorhees, 1998). The study confirmed that the comparative results for different runs remains stable despite changes in the underlying judgments. Taken together, these studies validate the use of the TREC collections for retrieval research.

4 EVALUATION

An important element of TREC is to provide a common evaluation forum. A standard evaluation pack-

```
<num> Number:  051

<dom> Domain:  International Economics

<title> Topic:  Airbus Subsidies

<desc> Description:
Document will discuss government assistance to Airbus Industrie, or mention
a trade dispute between Airbus and a U.S. aircraft producer over the issue of
subsidies.

<narr> Narrative:
A relevant document will cite or discuss assistance to Airbus Industrie by the
French, German, British or Spanish government(s), or will discuss a trade dispute
between Airbus or the European governments and a U.S. aircraft producer, most
likely Boeing Co.  or McDonnell Douglas Corp., or the U.S. government, over
federal subsidies to Airbus.

<con> Concept(s):
1.  Airbus Industrie
2.  European aircraft consortium, Messerschmitt-Boelkow-Blohm GmbH, British
Aerospace PLC, Aerospatiale, Construcciones Aeronauticas S.A.
3.  federal subsidies, government assistance, aid, loan, financing
4.  trade dispute, trade controversy, trade tension
5.  General Agreement on Tariffs and Trade (GATT) aircraft code
6.  Trade Policy Review Group (TPRG)
7.  complaint, objection
8.  retaliation, anti-dumping duty petition, countervailing duty petition,
sanctions
```

```
<num> Number:  168
<title> Topic:  Financing AMTRAK

<desc> Description:
A document will address the role of the Federal Government in financing the
operation of the National Railroad Transportation Corporation (AMTRAK).

<narr> Narrative:
A relevant document must provide information on the government's responsibility
to make AMTRAK an economically viable entity.  It could also discuss the
privatization of AMTRAK as an alternative to continuing government subsidies.
Documents comparing government subsidies given to air and bus transportation with
those provided to AMTRAK would also be relevant.
```

```
<num> Number:  207

<desc> What are the prospects of the Quebec separatists achieving independence
from the rest of Canada?
```

Figure 4: The evolution of TREC topic statements. Sample topic statement from TRECs 1 and 2 (top), TREC-3 (middle), and TREC-4 (bottom).

```
<num> Number:  312
<title> Hydroponics

<desc> Description:
Document will discuss the science of growing plants in water or some substance
other than soil.

<narr> Narrative:
A relevant document will contain specific information on the necessary nutrients,
experiments, types of substrates, and/or any other pertinent facts related to the
science of hydroponics.  Related information includes, but is not limited to, the
history of hydroponics, advantages over standard soil agricultural practices,
or the approach of suspending roots in a humid enclosure and spraying them
periodically with a nutrient solution to promote plant growth.
```

Figure 5: A sample TREC-6 topic.

	Min	Max	Mean
TREC-1 (51–100)	44	250	107.4
title	1	11	3.8
description	5	41	17.9
narrative	23	209	64.5
concepts	4	111	21.2
TREC-2 (101–150)	54	231	130.8
title	2	9	4.9
description	6	41	18.7
narrative	27	165	78.8
concepts	3	88	28.5
TREC-3 (151–200)	49	180	103.4
title	2	20	6.5
description	9	42	22.3
narrative	26	146	74.6
TREC-4 (201–250)	8	33	16.3
description	8	33	16.3
TREC-5 (251–300)	29	213	82.7
title	2	10	3.8
description	6	40	15.7
narrative	19	168	63.2
TREC-6 (301–350)	47	156	88.4
title	1	5	2.7
description	5	62	20.4
narrative	17	142	65.3
TREC-7 (351–400)	31	114	57.6
title	1	3	2.5
description	5	34	14.3
narrative	14	92	40.8

Table 5: Topic length statistics by topic section. Lengths count number of tokens in topic statement including stop words.

age, called `trec_eval`, is used to evaluate each of the submitted runs. `trec_eval` was developed by Chris Buckley at Cornell University and is available by anonymous ftp from `ftp.cs.cornell.edu` in the `pub/smart` directory. TREC reports a variety of recall- and precision-based evaluation measures for each run to give a broad picture of the run.

Since TREC-3 there has been a histogram for each system showing performance on each topic. In general, more emphasis has been placed in later TRECs on a "per topic analysis" in an effort to get beyond the problems of averaging across topics. Work has been done, however, to find statistical differences among the systems (see paper "A Statistical Analysis of the TREC-3 Data" by Jean Tague-Sutcliffe and James Blustein in the TREC-3 proceedings.) Additionally charts have been published in the proceedings that consolidate information provided by the systems describing features and system timing, allowing some primitive comparison of the amount of effort needed to produce the results.

Figure 4 shows two typical recall/precision curves. The x axis plots a fixed set of recall levels where

$$Recall = \frac{\textit{number of relevant items retrieved}}{\textit{total number of relevant items in the collection}}.$$

The y axis plots precision values at the given recall level, where precision is calculated by

$$Precision = \frac{\textit{number of relevant items retrieved}}{\textit{total number of items retrieved}}.$$

These curves represent averages over the 50 topics. The averaging method was developed many years ago (Salton & McGill, 1983) and is well accepted by the information retrieval community. The curves

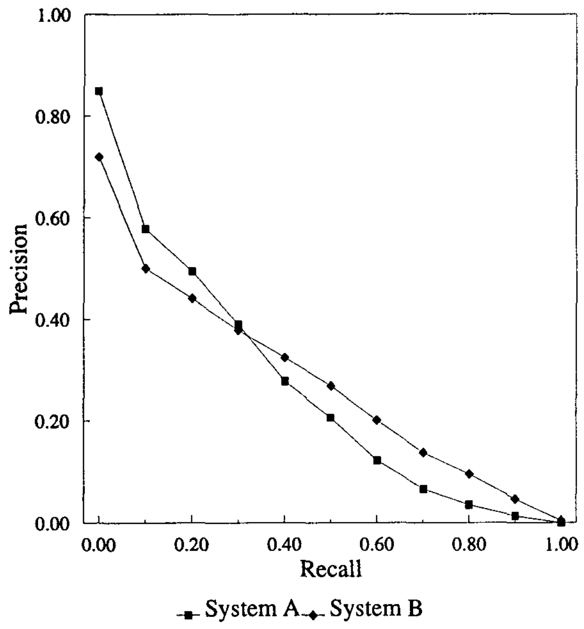

Figure 6: A sample Recall-Precision graph.

show system performance across the full range of retrieval, i.e., at the early stage of retrieval where the highly-ranked documents give high accuracy or precision, and at the final stage of retrieval where there is usually a low accuracy, but more complete retrieval. The use of these curves assumes a ranked output from a system. Systems that provide an unranked set of documents are known to be less effective and therefore were not tested in the TREC program.

The curves in Figure 4 show that system A has a much higher precision at the low recall end of the graph and therefore is more accurate. System B however has higher precision at the high recall end of the curve and therefore will give a more complete set of relevant documents, assuming that the user is willing to look further in the ranked list.

The single-valued evaluation measure most frequently used in TREC is the mean (non-interpolated) average precision. The average precision for a single topic is the mean of the precision obtained after each relevant document is retrieved (using zero as the precision for relevant documents that are not retrieved). The mean average precision for a run consisting of multiple topics is the mean of the average precision scores of each of the individual topics in the run. The average precision measure has a recall component in that it reflects the performance of a retrieval run across all relevant documents, and a precision component in that it weights documents retrieved earlier more heavily than documents retrieved later. Geometrically, mean average precision is the area underneath a non-interpolated recall-precision curve.

5 RETRIEVAL RESULTS

One of the important goals of the TREC conferences is that the participating groups freely devise their own experiments within the TREC task(s). For some groups, particularly the groups new to TREC, this means doing the ad hoc and/or routing task with the goal of achieving high retrieval effectiveness performance. Other groups use TREC as an opportunity to run experiments especially tuned to their own environment, either taking part in the organized tracks or performing associated tasks that can be evaluated easily within the TREC framework. The experimental work performed for TRECs 5, 6, and 7 is therefore both too broad and too extensive to be summarized within this paper. What is presented is some analysis of the trends within the ad hoc and routing tasks, plus a summary of the various tracks that have been run in these three TRECs. In all cases, readers are referred to the full TREC proceedings for papers from the various groups that give more details of their experiments.

5.1 The Ad Hoc Results

The basic TREC ad hoc paradigm has presented three major challenges to search engine technology from the beginning. The first is the vast scale-up in terms of number of documents to be searched, from several megabytes of documents to 2 gigabytes of documents. This system engineering problem occupied most systems in TREC-1, and has continued to be the initial work for most new groups entering TREC. The second challenge is that these documents are mostly full-text and therefore much longer than most algorithms in TREC-1 were designed to handle. The document length issue has resulted in major changes to the basic term weighting algorithms, starting in TREC-2. The third challenge has been the idea that a test question or topic contains multiple fields, each representing either facets of a user's question or the various lengths of text that question could be represented in. The particular fields, and the lengths of these fields, have changed across the various TRECs, resulting in different research issues as the basic environment has changed.

Because TREC-1 required significant system rebuilding by most participating groups due to the huge increase in the size of the document collection, the

TREC-1 results should be viewed as only very preliminary due to severe time constraints. TREC-2 occurred in August of 1993, less than 10 months after the first conference, and the TREC-2 results can be seen as both a validation of the earlier experiments on the smaller test collections and as an excellent baseline for the more complex experimentation that has taken part in later TRECs.

Table 5.1 summarizes the ad hoc task across the 6 TRECs that have occurred since 1992. It illustrates some of the common issues that have affected all groups, and also shows the initial use and subsequent spread of some of the now-standard techniques that have emerged from TREC.

Five different research areas are shown in the table, with research in many of these areas triggered by changes in the TREC evaluation environment. For example, the use of subdocuments or passages was caused by the initial difficulties in handling full text documents, particularly excessively long ones. The use of better term weighting, including correct length normalization procedures, made this technique less used in TREC's 4 and 5, but it resurfaced in TREC-6 to facilitate better input to relevance feedback.

The first research area shown in the table is that of term weighting. Most of the initial participants in TREC used term weighting that had been developed and tested on very small test collections with short documents (abstracts). Many of these algorithms were modified to handle longer documents in simple ways, however some algorithms were not amenable to this approach, resulting in some new fundamental research. The group from the Okapi system, City University, London (Robertson, Walker, Hancock-Beaulieu, & Gatford, 1994) decided to experiment with a completely new term weighting algorithm that was both theoretically and practically based on term distribution within longer documents. By TREC-3 this algorithm had been "perfected" into the BM25 algorithm now in use by many of the systems in TRECs 5, 6 and 7. Continuing along this same row in table 5.1, three other systems (the SMART system from Cornell (Singhal, Buckley, & Mitra, 1996), the PIRCS system from CUNY (Kwok, 1996) and the INQUERY system from the University of Massachusetts (Allan, Ballesteros, Callan, Croft, & Lu, 1996) changed their weighting algorithms in TREC-4 based on analysis comparing their old algorithms to the new BM25 algorithm. By TREC-5 many of the groups had adopted these new weighting algorithms, with the early adopters being those systems with similar structural models.

TREC-6 saw even further expansion of the use of these new weighting algorithms (alternatively called the Okapi/SMART algorithm, or the Cornell implementation of the Okapi algorithm). In particular, many groups adapted these algorithms to new models, often involving considerable experimentation to find the correct fit. For example IRIT (Boughanem & Soulé-Dupuy, 1998) modified the Okapi algorithm to fit a spreading activation model, IBM (Brown & Chong, 1998) modified it to deal with unigrams and trigrams, and the Australian National University (Hawking, Thistlewaite, & Craswell, 1998) and the University of Waterloo (Cormack, Clarke, Palmer, & To, 1998) used it in conjunction with various types of proximity measures. Of major note is the fact that City University also ran major experiments (Walker, Robertson, Boughanem, Jones, & Sparck Jones, 1998) with the BM25 weighting algorithm in TREC-6, including extensive exploration of the various existing parameters, and addition of some new ones involving the use of non-relevant documents!

It could be expected that 6 years of term weighting experiments would lead to a convergence of the algorithms. However, a snapshot of the top 8 systems in TREC-7 (see Table 5.1) shows that these systems are derived from many models and use different term weighting algorithms and similarity measures. Of particular note here is that new models and term weighting algorithms are still being developed, and that these are competitive with the more established methods. This applies both to new variations on old weighting algorithms, such as the double log tf weighting from AT&T (Singhal, Choi, Hindle, Lewis, & Pereira, 1999) and to more major variations such as the new weighting algorithm from TNO (Hiemstra & Kraaij, 1999), and the completely new retrieval model from BBN (Miller, Leek, & Schwartz, 1999).

The second new technique started back in TREC-2 (the second line of table 5.1) was the use of smaller sections of documents, called subdocuments, by the PIRCS system at City University of New York (Kwok & Grunfeld, 1994). Again this issue was forced by the difficulty of using the PIRCS spreading activation model for documents having a wide variety of lengths. By TREC-3 many of the groups were also using subdocuments, or passages, to help with retrieval. But, as mentioned before, TREC's 4 and 5 saw far less use of this technique as many groups dropped the use of passages due to minimal added improvements in performance.

TREC-6 saw a revival in the use of passages, but generally only for specific uses. Whereas the PIRCS system continued to use 550-word subdocu-

	TREC-2	TREC-3	TREC-4	TREC-5	TREC-6	TREC-7
Term weighting	baseline for most systems beginning of Okapi weighting experiments	Okapi perfects BM25 algorithm	new weighting algorithms in SMART, INQUERY, and PIRCS systems	use of Okapi / SMART weighting algorithms by other groups	adaptations of Okapi / SMART algorithms in most systems	new retrieval models by TNO and BBN
Passages	use of subdocuments by PIRCS system	heavy use of passages / subdocuments	decline in use of passages		use of passages in relevance feedback	multiple uses of passages
Automatic query expansion		beginning of expansion using top X documents	heavy use of expansion using top X documents	beginning of more complex expansion schemes	more sophisticated expansion experiments by many groups	
Manual query modification		beginning of manual expansion using other sources	major experiments in manual editing, user-in-the-loop	extensive user-in-the-loop experiments	simpler user-specific strategies tested	
Other new areas		initial use of "data fusion"		start of more concentration on initial topic	more complex use of data fusion continued focus on initial topic, especially the title	

Table 6: Use of new techniques in the ad hoc task

Organization	Model	Weighting/Similarity	Phrase Imp.	Comments
Okapi group	probabilistic	BM25	minimal*	*last reported in TREC-5
AT&T Labs Research	vector	pivot*		*byte normalization
U. Mass	inference net	belief function	3.6%	
RMIT/UM/CSIRO	vector	BM25/cosine		phrases used
BBN	HMM	probabilistic	2%	bigram phrases
TwentyOne	vector	new probabilistic		no phrases used
CUNY	spread. act.	avtf/RSV	2%	phrases used for reranking
Cornell/SabIR	vector	pivot		

Table 7: Models and term weight in TREC-7.

ments for all its processing, most systems used passages only in the topic expansion phase. The Australian National University (Hawking et al., 1998) worked with "hot spots" of 500 characters surrounding the original topic terms to locate new expansion terms. AT&T (Singhal, 1998) used overlapping windows of 50 words to help rerank the top 50 documents before selecting the final documents for use in expansion. The University of Waterloo (Cormack et al., 1998) used passages of maximum length 64 words to select expansion terms, whereas Verity (Pedersen, Silverstein, & Vogt, 1998) used their automatic summarizer for this purpose. Two groups (Lexis-Nexis (Lu, Meier, Rao, Miller, & Pliske, 1998) and MDS (Fuller et al., 1998)) performed major experiments in the use of passages, particularly when employed in conjunction with other methods as input to data fusion. This diverse use of passages continued in TREC-7, with passages clearly becoming one of the standard tools for experimentation.

The query expansion/modification techniques shown in the third and fourth lines of the table 5.1 were started when the topics were substantially shortened in TREC-3. As described in section 3.2, the format of the topics was modified to remove a valuable source of keywords: the concept section. In the search for some technique that would automatically expand the topic, several groups revived an old technique of assuming that the top retrieved documents are relevant, and then using them in relevance feedback. This technique, which had not worked on smaller collections, turned out to work very well in the TREC environment.

By TREC-6 almost all groups were using variations on expanding queries using information from the top retrieved documents (often called *pseudo-relevance feedback*). There are many parameters needed for success here, such as how many top documents to use for mining terms, how many terms to select, and how to weight those terms. There has been general convergence on some of these parameters. Table 5.1 shows the characteristics of the expansion tools used in the top 8 systems in TREC-7. The second column gives the basic expansion model, with the vector-based systems using the Rocchio expansion and other systems using expansion models more suitable to their retrieval model. For example, the Local Context Analysis (LCA) method developed by the INQUERY group (Xu & Croft, 1996) has been successfully used by other groups. The third column shows the number of top-ranked documents (P if passages were used), and the number of terms added from these documents. It should be noted that these numbers are more similar than in earlier TRECs, although they are still being investigated by new systems adopting these techniques as there can be subtle differences between systems that strongly influence parameter selection. The fourth column shows the source of the documents being mined for terms, which has generally moved to the use of as much information as possible, i.e. all the TREC disks as opposed to only those being used for testing purposes.

TRECs 5, 6, and 7 saw many additional experiments in the query expansion area. The Open Text Corporation (Fitzpatrick & Dent, 1997) gathered terms for expansion by looking at relevant documents from past topics that were loosely similar to the TREC-5 topics. Several groups ((Lu, Ayoub, & Dong, 1997; Namba, Igata, Horai, Nitta, & Matsui, 1999)) have tried clustering the top retrieved documents in order to more accurately select expansion terms, and in TREC-6 three groups (City University, AT&T, and IRIT) successfully got information from negative feedback, i.e. using non-relevant documents to modify the expansion process.

TREC-7 contained even more experiments in automatic query expansion, such as the group (Mandala, Tokunaga, Tanaka, Okumura, & Satoh, 1999) that compared the use of three different thesauri for expansion (WordNet, a simple co-occurrance the-

Organization	Expansion/Feedback	Top Docs/Terms added	Disks used	Comments
Okapi group	probabilistic	Full–15/30 T+D–10/30 T only–6/20+title	1-5	
AT&T Labs Research	Rocchio	10/20+5 phrases	1-5	conservative enrichment
U. Mass	LCA	30P/50	1-5	reranking using title terms before expansion
RMIT/UM/CSIRO	Rocchio	10/40+5 phrases	?	additional experiments with passages
BBN	HMM-based	6/?	?	differential weighting on topic parts
TwentyOne	Rocchio	3/200	?	
CUNY	LCA	200P/?	1-5	
Cornell/SabIR	Rocchio	30/25	4-5	clustering, reranking

Table 8: Characterization of query expansion used in best automatic ad hoc TREC-7 runs.

saurus and an automatically built thesaurus using predicate-argument structures). Of particular note is the AT&T (Singhal et al., 1999) investigation into "conservative enrichment" to avoid the additional noise caused by using larger corpora (all five disks) for query expansion.

Groups that build their queries manually also looked into better query expansion techniques starting in TREC-3 (see fourth line of table 5.1). At first these expansions involved using other sources to manually expand the initial query. However the rules governing manual query building changed in TREC-5 to allow unrestricted interactions with the systems. This change caused a major evolution in the manual query expansion, with most systems not only manually expanding the initial queries, but then looking at retrieved documents in order to further expand the queries, much in the manner that users of these systems could operate. Two types of experiments were notable in TREC-5: those that could be labelled as "manual exploration" runs and those that involved a more complex type of human-machine interaction. The first type is exemplified by the GE group (Strzalkowski et al., 1997), where the task was to ask users to pick out phrases and sentences from the retrieved documents to add to the query, in hopes that this process could be imitated by automatic methods. The CLARITECH group (Milic-Frayling, Evans, Tong, & Zhai, 1997) is a good example of the second type of manual TREC-5 runs. They examined a multi-stage process of query construction, where the goal was to investigate better sets of tools that allow users to improve their queries, including different sources for suggestions of expansion terms and also various levels of user-added constraints to the expansion process.

Many of the manual experiments seen in both TREC-6 and TREC-7, however, hark back to the simpler scenario of having users edit the automatically-generated query, or having users select documents to be used in automatic relevance feedback. Several of the groups had specific user strategies that they tested.

- GE Corporate R&D/Rutgers University (Strzalkowski, Lin, & Perez-Carballo, 1998) used automatically-generated summaries of the top 30 documents retrieved as sources of manually-selected terms and phrases.
- CLARITECH Corp. (Evans, Huettner, Tong, Jansen, & Bennett, 1999) performed a user experiment measuring the difference in performance between two presentation modes: a ranked list vs a clustered set of documents.
- University of Toronto (Bodner & Chignell, 1999) used their dynamic hypertext model to build the queries.
- Lexis-Nexis (Rao, Humphrey, Parhizgar, Wilson, & Pliske, 1999) experimented with human rele-

vance feedback as opposed to automatic feedback from the top 20 documents.

The final line in table 5.1 shows some of the other areas that have seen concentrated research in the ad hoc task. Data fusion has been used in TREC by many groups in various ways, but has increased in complexity over the years. For example, a project involving four teams led by Tomek Strzalkowski has continued the investigation of merging results from multiple streams of input using different indexing methods ((Strzalkowski et al., 1997, 1998, 1999). In TREC-6, several groups such as Lexis-Nexis (Lu et al., 1998) and MDS (Fuller et al., 1998) used multiple stages of data fusion, including merging results from different term weighting schemes, various mixtures of documents and passages, and different query expansion schemes.

The INQUERY system from the University of Massachusetts has worked in all TREC's to automatically build more structure into their queries, based on information they have "mined" from the topics (Brown, 1995). Starting in TREC-5, there have been experiments by other groups to use more information from the initial topic. Lexis-Nexis (Lu et al., 1997) used the inter-term distance between nouns in the topic. Several other groups have made use of term proximity features (Australian National University (Hawking, Thistlewaite, & Bailey, 1997), University of Waterloo (Clarke & Cormack, 1997) , and IBM) to improve retrieval scores, while others (CUNY (Kwok & Grunfeld, 1997), AT&T (Singhal, 1998), and INQUERY (Allan, Callan, Sanderson, Xu, & Wegmann, 1999)) have used the initial topic to look for clues that would suggest a need for more emphasis on certain topic terms. TREC-7 had two additional groups working with the use of term cooccurrance and proximity as alternative methods for ranking (see (Braschler, Wechsler, Mateev, Mittendorf, & Schäuble, 1999) and (Nakajima, Takaki, Hirao, & Kitauchi, 1999)).

A final theme that has continued throughout all the TREC conferences has been the investigation of the use of phrases in addition to single terms. This has long been a topic for research in the information retrieval community, with generally unsuccessful results. However there was initially hope that use of phrases in these much larger collections would become critical and almost all groups have experimented with phrases. In general these experiments have been equally unsuccessful.

The fourth column of table 5.1 shows the widespread use of phrases in addition to single terms in TREC-7, but the minimal improvement from their use. The biggest improvement reported in the papers was 3.6% from the INQUERY group at the University of Massachusetts (Allan et al., 1999). Whereas most of the other groups are also using phrases, many did not bother to test for differences due to minimal results in earlier years. Cornell/SabIR reported 7.7% improvement in TREC-6, but this is the improvement on top of the initial baseline, not the improvement after expansion. Private conversations with several of these groups indicate that these improvements are likely to be much less if measured after expansion. As is often the case, these minimal changes in the averages cover a wide variation in phrase performance across topics. A special run by the Okapi group (many thanks) showed less than a 1% average difference in performance, but 19 topics helped by phrases, 14 hurt, and the rest unchanged. Whereas the benefit of phrases is not proven, they are likely to remain a permanent tool in the retrieval systems in a manner similar to the earlier adoption of stemming.

It is interesting to note that many of these groups are using different phrase "gathering" techniques. The Okapi group has a manually-built phrase list with synonym classes that has slowly grown over the years based on mostly past TREC topics. The automatically-produced INQUERY phrase list was new for TREC-6 (Allan et al., 1998), the Cornell list was basically unchanged from early TRECs, and the BBN list was based on a new bigram model.

The creation of two formal topic lengths in TREC-5 has inspired many experiments comparing results using those different topic lengths, and the addition of a formal "title" in TREC-6 increased these investigations. Table 5.1 shows the results (official and unofficial as reported in the papers) of the top 8 TREC-7 groups showing their use of different topic parts. The second column gives the various topic parts used by each group (T = title, D = description, N = narrative). The third column gives the average precision using only the description and title. The fourth and fifth columns give the corresponding performance of the systems using either only the title or using the full topic (all topic parts).

Note that most of the best runs use the full topic. However there is now a smaller performance difference between runs that use the full topic and runs that use only the title and description sections than was seen in earlier TRECs. This is most likely due to improved query expansion methods, but could be due to variations across topic sets. It should be noted that the improvement going to the full topic is only 1% for several groups. The decrease in performance using only the title is more marked, ranging from 4%

	Long	Desc	Title
Okapi	28	13	9
CUNY	27	10	13
Cornell	22	17	11

Table 10: Number of TREC-7 topics performing best by topic length.

to 22%. The TREC-7 title results should be a truer measure of the effects of using the title only than TREC-6, where the descriptions were often missing key terms. However, it is not clear how representative these titles are with respect to very short user inputs and therefore title results should best be viewed as how well these systems could perform on very short, but very good user input.

Looking at individual topic results shows a less consistent picture. Table 5.1 shows the number of topics that had the best performance from among a group's three runs using different input lengths. Not only is there a wide variation across topics, there is also a wide variation across systems in that topics that work best at a particular length for one group did not necessarily work best at that length for the other groups.

5.2 The Routing Results

The routing evaluation used a specifically selected subset of the training topics against a new set of test documents, but there have always been difficulties in locating appropriate testing data for the routing task. TREC-3 was forced to re-use some of the training data, and TREC-4 performed routing tests using the *Federal Register* (with new data) for 25 of the topics, and using training data and "net trash" for testing the other 25 topics. This situation was clearly not ideal and for TREC-5 NIST held back decisions on the routing topics until a new data source could be found.

When the FBIS data became available, it was decided to pick topics that had many relevant documents in the *Associated Press* data, on the assumption that the FBIS data would be similar to AP. Because of delays in getting and processing the data, this assumption could not be checked out, and problems arose that will be discussed later.

It should be noted that the routing task in TREC has always served two purposes. The first is its intended purpose: to test systems in their abilities to use training data to build effective filters or profiles. The second purpose, which has become equally important in the more recent TRECs, is to serve as a learning environment for more effective retrieval techniques in general. Groups use the relevance judgments to explore the characteristics of relevant documents, such as which features are most effective to use for retrieval or how to best merge results from multiple queries. This is more profitable than simply using the previous TREC results in a retrospective manner because of the use of completely new testing data for evaluation.

A focus on using the training data as a learning environment was particularly prevalent in TREC-5. Cornell (Buckley, Singhal, & Mitra, 1997) used the relevant and non-relevant documents for investigations of Rocchio feedback algorithms, including more complex processes of expansion and weighting. The University of Waterloo (Clarke & Cormack, 1997) interactively searched the training data for co-occurring substrings and GE (Strzalkowski et al., 1997) ran major experiments in data fusion to test their new stream-based architecture. In each of these cases the experiments are assumed to lead to better ways of doing the routing task, and also to new approaches for the ad hoc task.

Three experimental themes dominate most routing experiments in TREC-5. The first is the discovery of optimal features (usually single terms) for use in the query or filter. The Okapi system from City University, London (Beaulieu et al., 1997) continued its experiments in repeatedly trying various combinations of terms to discover the optimal set, but for TREC-5 used subsets of the training data. The University of California at Berkeley (Gey, Chen, He, Xu, & Meggs, 1997) concentrated on further investigations of the use of the chi-square discrimination measure to locate large numbers of good terms, and the Swiss Federal Institute of Techology (ETH) (Ballerini et al., 1997) tried three different feature selection methods, including the chi-square method, the RSV (OKAPI) method, and a new method, the U measure. Xerox (Hull et al., 1997) also investigated a new feature selection method, the binomial likelihood ratio test.

The second theme was the use of co-occurring term pairs in the training data to "expand" the query. Four groups experimented with locating and incorporating co-occurring pairs of terms, including the INQUERY group from the University of Massachusetts in both TREC-4 and TREC-5 (Allan et al., 1996, 1997), and Cornell University in TREC-5 (Buckley et al., 1997). As mentioned before, Waterloo interactively looked for word-pairs or co-occurring strings to manually add to their query. ETH used the OKAPI RSV values to formally motivate a series of experi-

Organization	Topic Parts	D + T	T only	Full Topic	Comments
Okapi group	T,D,N	0.281	0.253 (-10%)	0.284 (1%)	fused run-0.296
AT&T Labs Research	T,D	0.296	0.249 (-16%)		
U. Mass	T,D,N	0.252		0.274 (9%)	title filtered run-0.282
RMIT/UM/CSIRO	T,D	0.281	0.220 (-22%)	0.285 (1%)	
BBN	T,D,N			0.280	
TwentyOne	T,D,N			0.279	
CUNY	T,D,N	0.254	0.243 (-4%)	0.266 (5%)	with phrases-0.272
Cornell/SabIR	T,D,N	0.254*	0.239 (-6%)	0.267 (5%)	*description only

Table 9: TREC-7 Performance using variations in topic length.

ments using co-occurring terms within different portions of the document (within sentence, within paragraph, etc.) as different methods of constructing queries. These multiple representations of the query were then linearly combined, with the parameters for that combination discovered using logistic regression on the training data.

The third theme in the routing experiments was the continuing effort to use only subsets of the training data. The number of judged documents per topic is on the order of 2000 or more, and this can be computationally difficult for complex techniques. Efficiency has motivated CUNY experiments (the PIRCS system) since TREC-3 where they tried using only the "short" documents for training. In TREC-5 this group (Kwok & Grunfeld, 1997) used genetic algorithms to select the optimal set of training documents. Cornell (in TREC-5) used a new "query zone" technique to subset the training documents so that not all non-relevant documents were used for training. The goal was not just improved efficiency, but also improved effectiveness in that training was more concentrated on documents that the Cornell system was likely to retrieve.

There is another issue that suggests the use of subsets: the problem of overfitting the queries/methods to the training data. This was specifically emphasized in the City system, where they used different subsets of the training data for locating features, and used combinations of runs for their final results. Xerox used subsets to reduce overfitting, with their subsets based on finding documents within a "local zone" to the query (a predecessor to the query zoning technique used by Cornell). The Xerox paper provides more discussion of the overfitting problem and suggests some additional techniques to avoid it.

As in the ad hoc task, there is a heavy adoption rate across groups for successful techniques. For the ad hoc task these techniques revolve around better ways of handling the initial topic, or use of the top X documents for relevance feedback. Because of the existence of training data in routing, the routing experiments have generally not used the topic itself heavily, but constructed queries mainly based on the training data. The success of these techniques therefore revolves around how well the test data matches the training data, and also on how tuned the techniques are to the particular training data.

TREC-5 used AP documents as training data, with FBIS material for test data. Whereas the types of documents are similar, the domains of the documents did not always match. For some topics there was a good match of training and test data, but for others the match was very poor, and very few relevant documents were found for those topics. Four topics had zero relevant documents in the test set, and an additional six topics had only one or two relevant documents. Additionally there was a serious mismatch on the number of relevant documents for a topic in the training data and in the test data. Even after dropping the four topics with no relevant documents from the evaluation, the results are still heavily affected by the mismatch. The overall results for TREC-5 were not better than for TREC-4 (or TREC-3.

In TREC-6 an attempt was made to have a close match between the training and test data. Since the TREC-5 routing task had used a document stream from the Foreign Broadcast Information Service (FBIS) as its test set, a new stream of FBIS documents was selected as the TREC-6 test set. The TREC-6 routing topics consisted of 38 topics used in TREC-5 that had at least 5 relevant documents in the original FBIS stream, plus nine new topics (that had minimal training data on the original FBIS stream). The histogram in Figure 7 shows that the training and test data do have similar numbers of relevant documents for most topics.

The following gives the various experiments that were run by the 8 top performing systems in the TREC-6 routing task.

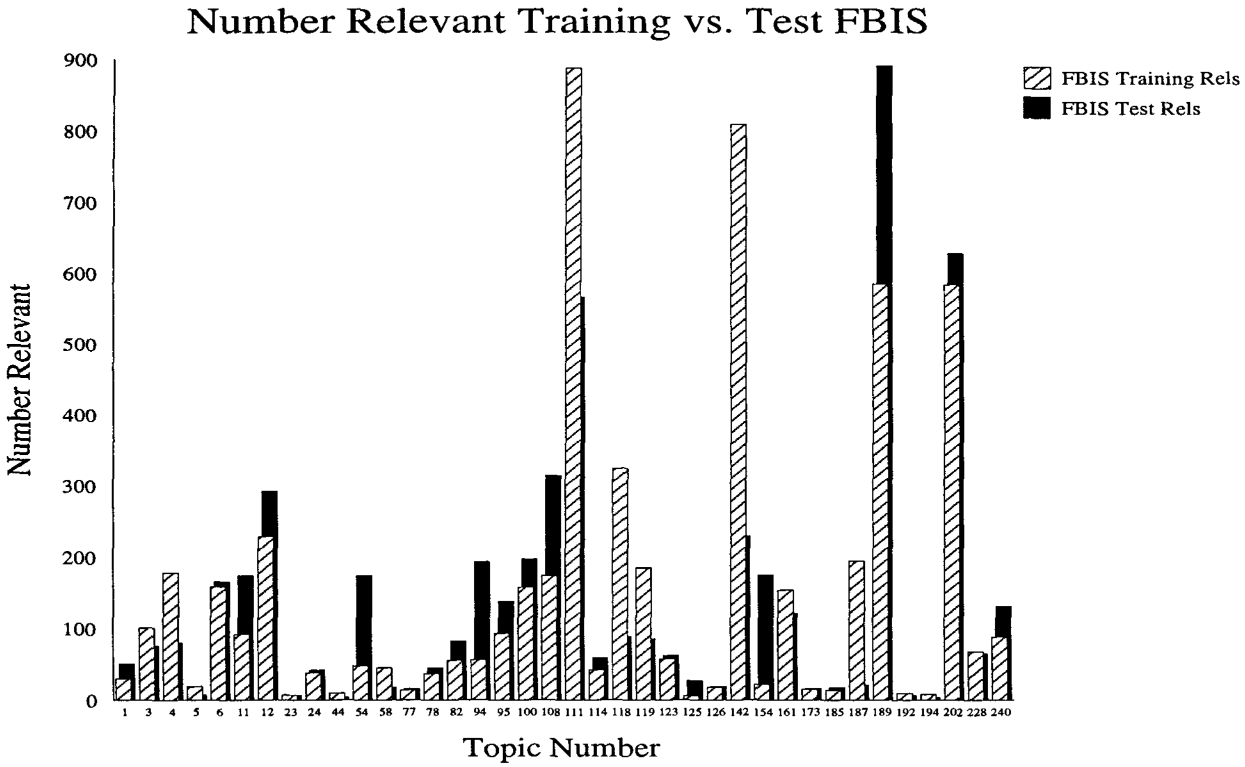

Figure 7: Comparison of the number of relevant documents in the training and test FBIS collections.

- AT&T Labs Research (Singhal, 1998) added the machine learning technique of boosting to the query refinement phase of the Cornell TREC-5 routing algorithm (which includes the use of word pairs, DFO optimization, and query zones).
- City University, London (Walker et al., 1998) explored iterative methods of term weighting with the goal of avoiding overfitting.
- Cornell/SaBIR Research (Buckley, Mitra, Walz, & Cardie, 1998) also used a variant of the basic Cornell TREC-5 routing approach, adding SuperConcepts to the routing query.
- Queens College, CUNY (Kwok, Grunfeld, & Xu, 1998) combined results from five separate component runs; this combined result is superior to each of the individual components.
- University of Waterloo (Cormack et al., 1998) interactively refined a set of Boolean queries into a single tiered Boolean query for each topic.
- Claritech Corporation (Milic-Frayling, Zhai, Tong, Jansen, & Evans, 1998) explored the benefits of using different term selection methods in different parts of the query refinement process. For this run they developed different queries using different term selection strategies and then, for each topic, selected the query that performed the best on the training data.
- MSI/IRIT/SIG/CERISS (Boughanem & Soulé-Dupuy, 1998) continued their work with a spreading activation model by expanding queries with the top 30 terms from relevance backpropagation.
- Swiss Federal Institute of Technology (ETH) (Mateev, Munteanu, Sheridan, Wechsler, & Schäuble, 1998) also performed a combination run where one component run selected query words and phrases based on the U-measure.

The best mean average precision for a routing run in TREC-6 was .420, a 9% improvement over TREC-5's best of .386. However, given that the TREC-6 task was designed to use a homogeneous data set whereas the TREC-5 test data were different from the training data, a greater improvement was expected. At this point, it is unclear why the difference was not greater. It is possible that while the numbers of relevant documents in the training and test set are comparable, the relevant documents in each set don't "look like" each other. However, this is unlikely since both sets of documents come from a common source. It is also possible that the mismatch between training and test sets is not as significant a factor as was thought.

Another hypothesis suggested by (Singhal, 1998) is that the relevance judgments are less consistent for routing than they are for the ad hoc task, and that this inconsistency prevents the machine learning methods that are prevalent in the task from performing well. Since some routing topics have been used many times, and therefore have relevance judgments spanning many years, the judgments are likely to be less consistent than for the ad hoc task. It may be instructive to explore the stability of the routing techniques in the face of different relevance judgments, especially given that real user judgments are known to be extremely volatile (Schamber, 1994).

Because of operational constraints on the overall TREC program, it was decided to pursue further investigations in routing within the very closely related filtering track. For this reason there was no routing task in TREC-7, but there was a routing option in the filtering track.

6 THE TRACKS

One of the goals of TREC is to provide a common task evaluation that allows cross-system comparisons, and this has proven to be a key strength in TREC. A second major strength is the loose definition of the ad hoc task, which allows a wide range of experiments. The addition of secondary tasks (called tracks) in TREC-4 combined these strengths by creating a common evaluation for retrieval subproblems. TREC participants are free to turn in results for any, or all, or none, of the tracks.

The tracks have had a significant impact on TREC participation. Figure 8 shows the number of experiments performed in each TREC, where the set of runs submitted for a track by one group is counted as one experiment. The number of experiments increased each year through TREC-6 then decreased in TREC-7, mostly due to the elimination of the routing main task and the Chinese track. The number of participants performing the ad hoc task continues to grow, with 42 groups taking part in TREC-7 compared to 31 in TREC-6. The number of participants in each of the TREC-7 tracks and the corresponding TREC-6 participation is given below.

	TREC-6	TREC-7
CLIR	13	9
filtering	10	12
HP	5	4
interactive	9	8
query	0	2
SDR	13	10
VLC	7	6

The set of tracks run in any particular year depends on the interests of the participants and sponsors, as well as on the suitability of the problem to the TREC environment. Some initial tracks have been discontinued because the goals of the track were met. For example, the Spanish track, an ad hoc task in which both topics and documents are in Spanish, was discontinued when the results demonstrated that current retrieval systems can retrieve Spanish documents as effectively as English documents. Other tracks, such as the interactive track, have been run each year, but have changed their focus in different years. Each track has a set of guidelines developed under the direction of the track coordinator. The set of tracks and their primary goals are listed below. See the track reports in the various TREC proceedings for a more complete description of each track and its results.

6.1 The Spanish and Chinese Tracks

Track reports– (Smeaton & Wilkinson, 1997; Wilkinson, 1998)

The first non-English track was started in TREC-3. Four groups worked with 25 topics in Spanish, using a document collection consisting of about 200 megabytes (58,000 documents) of a Mexican newspaper from Monterey (*El Norte*). Since there was no training data for testing (similar to the startup problems for TREC-1), the groups used simple techniques. The major result from this very preliminary experiment in a second language was the ease of porting the retrieval techniques across languages. Cornell (Buckley, Salton, Allan, & Singhal, 1995) reported that only 5 to 6 hours of system changes were necessary (beyond creation of any stemmers or stopword lists).

In TREC-4 10 groups took part, using the same document collection and 25 new topics. The final round of Spanish retrieval took place in TREC-5, again with 25 new topics and also with additional text (1994 newswire from *Agence France Presse*, including 308 megabytes or 173,950 documents). Seven groups took part in Spanish, with several of them building more elaborate procedures for testing, such

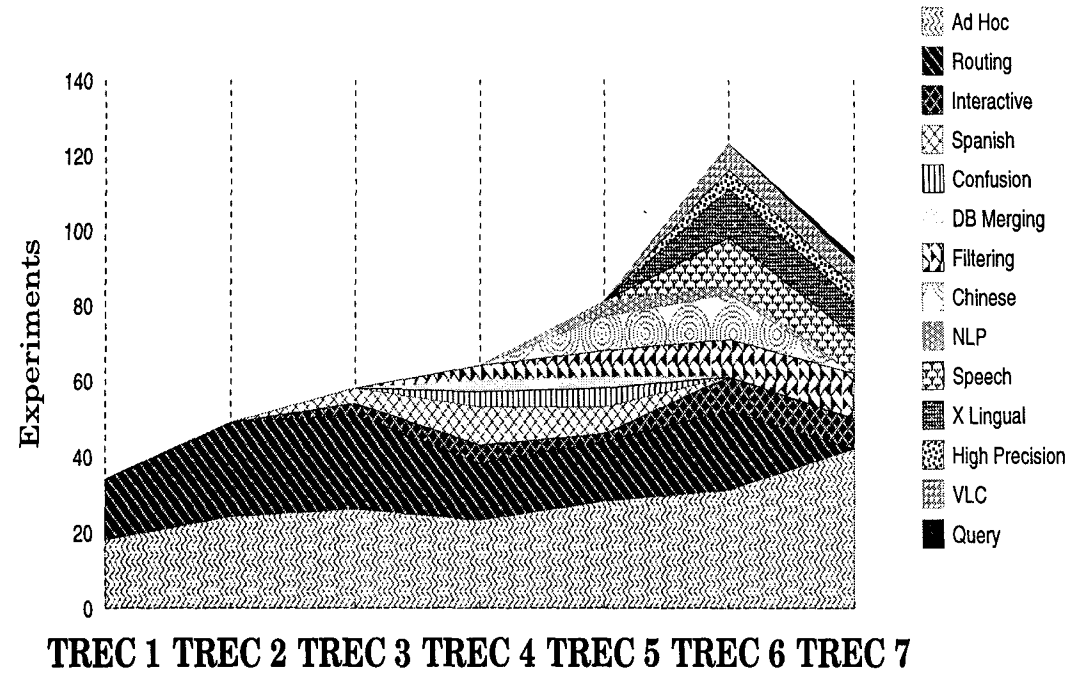

Figure 8: Number of TREC experiments by TREC task

as Spanish POS taggers. But in the main these did not improve performance and the major outcome of the Spanish track was that most of the techniques used in English retrieval, including the advanced ones used in the ad hoc task, can be successfully applied to Spanish.

The purpose of the Chinese track was to investigate retrieval performance for a language whose orthographics are not word-oriented. Participants performed an ad hoc search in which both the topics and the documents were in Chinese. The document set was a collection of articles selected from the *Peoples Daily* newspaper and the *Xinhua* newswire, a total of 168,811 documents in 170 megabytes. Twenty-eight topics were created for the track in TREC-5 and an additional 26 topics for TREC-6.

Nine groups submitted Chinese runs in TREC-5, and since it was the first year for Chinese in TREC, most groups concentrated on segmentation issues. In TREC-6 there were 12 participating groups, and again the majority of the experiments compared different methods of segmenting the text into retrieval features. In general, approaches that used single characters or bi-grams as features were competitive with word-based approaches and had the advantage of not requiring complicated segmentation schemes.

A confounding factor in the analysis of the retrieval results was that the retrieval effectiveness was quite high (the median mean average precision was greater than 0.5), and was similar across systems. It was difficult to distinguish more effective techniques when all techniques appear to work equally as well. Without more testing, it was not possible to determine whether the TREC-6 topics were simply easy, or if there is something inherent in Chinese that facilitates retrieval. Further testing was postponed until new Chinese data could be assembled.

6.2 The Cross Language (CLIR) Track

Track reports– (Schäuble & Sheridan, 1998; Braschler, Krause, Peters, & Schäuble, 1999)

The CLIR task focuses on searching for documents in one language using topics in a different language. The first CLIR track was held in TREC-6 (Schäuble & Sheridan, 1998). Three document sets were used: a 250 MB set of French documents from the Swiss news agency *Schweizerische Depeschen Agentur* (SDA); a 330 MB set of German documents from SDA plus a set of articles (200 MB) from the newspaper *New Zurich Newspaper* (NZZ); and a 750 mB set of English documents from the AP newswire. All of the document sets contain news stories from approximately the same time period, but are not aligned or specially coordinated with one another. A set of 25 topics that were translated into each of the languages was also provided. Participants searched for documents in one target language using topics written in a different language. In addition, participants were asked to perform a monolingual run in the target language to act as a baseline.

Thirteen groups participated in the TREC-6 CLIR track. Three major approaches to cross-language retrieval were represented: machine translation, where either the topics or the documents were translated into the target language; the use of machine-readable bilingual dictionaries or other existing linguistic resources; and the use of corpus resources to train or otherwise enable the cross-language retrieval mechanism. The approaches all behaved similarly in that some group obtained good cross-language performance for each method. In general, the best cross-language performance was between 50%–75% as effective as a quality monolingual run.

The TREC-7 task expanded on this beginning. The document set for the TREC-7 track consisted of all the documents used in the TREC-6 track plus the Italian version of the SDA for the same time period. Participants were provided with a new set of 28 topics (with translations available in English, French, German, and Italian), and used one topic language to search the combined document set. That is, a single run retrieved documents written in different languages. To enable participation in the track by more groups, a second task was also defined in which English topics were run against the combined French and English document set.

The TREC-7 track also defined an optional subtask. The subtask used a different document collection, a 31,000 document structured database (formatted as SGML fielded text data) from the field of social science plus the NZZ articles, and a separate set of 28 topics. The rational of the subtask was to study CLIR in a vertical domain (i.e. social science) where a German/English thesaurus is available.

Nine groups participated in the TREC-7 CLIR track, with five groups performing the test on the full four-language collection, and seven groups performing the test on the English and French collection. No runs were submitted for the optional subtask; however this subtask is planned to be repeated in TREC-8 now that groups have more experience with cross language retrieval. The results of the track demonstrate that very different approaches to cross-language retrieval can lead to comparable retrieval effectiveness.

The construction of the cross language test collection differs from the way the other TREC collections have been created. The set of topics created for the track were developed at four different institutions: NIST (English); EPFL Lausanne, Switzerland (French); University Bonn, Germany (German); and CNR, Pisa, Italy (Italian). Each institution created topics that would target documents in their corresponding language. The relevance judgments for all topics for a particular document language were also made at the site responsible for that language. This is the first time that TREC has used multiple relevance assessors for a single topic.

6.3 The Filtering Track

Track reports– (Lewis, 1997; Hull, 1998, 1999)

As mentioned before, the routing task investigates the performance of systems that use standing queries to search new streams of documents. As the routing task is defined in TREC, participants use old topics with existing relevance judgments to form routing queries. These queries are then run against a previously unseen document collection to produce a ranked document list. However, real routing applications generally require a system to make a binary decision whether or not to retrieve the current document, not to form a ranking of a document set. The filtering track was started in TREC-4 to address this more difficult version of the routing task.

The question of how to evaluate filtering runs has been a focus of the filtering track since its inception. Since filtering results are an unordered set of documents, the rank-based measures used in the ad hoc and routing tasks are not appropriate. The main approach has been to try utility functions as measures of the quality of the retrieved set—the quality is computed as a function of the benefit of retrieving a relevant document and the cost of retrieving an irrelevant document.

In TREC-5, a family of three functions was tried in an investigation of how retrieval was affected by changes in the relative worth of retrieving a relevant document versus not retrieving a nonrelevant document. There were seven participating groups, but the major outcome was the awareness of the difficulty of defining an appropriate utility measure.

In TREC-6 two different utility functions were used:

$$\begin{aligned} \mathrm{F1} &= 3\mathrm{R}^+ - 2\mathrm{N}^+ \\ \mathrm{F2} &= 3\mathrm{R}^+ - \mathrm{N}^+ - \mathrm{R}^- \end{aligned}$$

where R^+ is the number of relevant documents that are retrieved, R^- is the number of relevant documents that are not retrieved, and N^+ is the number of non-relevant documents that are retrieved.

A problem with utilities as measures is that different topics have widely varying possible utility values, and these utilities cannot be normalized. Thus, utilities cannot be meaningfully averaged or compared across topics. A second measure, ***average set precision*** (ASP) defined as the product of recall and precision, was therefore introduced in TREC-6. Unfortunately, ASP suffers from its own drawback. When no relevant documents are retrieved, ASP is 0 regardless of how many non-relevant documents are retrieved. This is a problem in filtering evaluation since knowing when to NOT retrieve documents is an important part of the filtering task.

The F1 utility measure defined above rewards a system with 3 "points" for every relevant document document it retrieves, and penalizes the system two "points" for every nonrelevant document it retrieves. While these benefit and cost values seem reasonable, they define a level of performance that is quite challenging for current systems to meet. Ten groups participated in the TREC-6 filtering track and submitted a total of 17 runs that were optimized for the F1 measure. The best of these runs had a positive utility for 33 (of 47) topics, and the median of the 17 runs had a positive utility for just 20 topics. Since retrieving no documents has an F1 utility of 0, retrieving no documents would result in a better F1 utility than current systems obtain on average.

The F2 utility measure is even more demanding since systems are penalized for not retrieving relevant documents (and thus retrieving no documents also results in a negative utility). Of the 17 runs optimized for the F2 utility, only 10 topics had a positive median F2 utility.

The TREC-7 filtering track contained three tasks of increasing difficulty (and realism). For each task, topics 1–50 and the AP newswire collection on Disks 1–3 were used (with different splits into training and test sets, depending on the task). The first task was the traditional routing task. The second task was a ***batch*** filtering task in which systems are given topics and relevance judgments as in the routing task, and must then decide whether or not to retrieve each document in the test portion of the collection. This task is what previous filtering tracks in TRECs 5 and 6 had performed.

The third task, and the focus of the TREC-7 track, was an *adaptive* filtering task. In this task, a filtering system starts with just the query derived from the topic statement, and processes documents one at a time in date order. If the system decides to retrieve a document, it obtains the relevance judgment for it, and can modify its query as desired.

In TREC-7 two different utility functions were used:

$$\begin{aligned} \mathrm{F1} &= 3\mathrm{R}^+ - 2\mathrm{N}^+ \\ \mathrm{F3} &= 4\mathrm{R}^+ - \mathrm{N}^+ \end{aligned}$$

where R^+ and N^+ are the number of relevant and non-relevant documents retrieved, respectively. An approach to scaling and normalizing utilities was introduced in this year's track(Hull, 1999)

Twelve groups submitted at least one TREC-7 filtering run. A total of 46 runs were submitted, consisting of 10 routing runs, 12 batch filtering runs, and 24 adaptive filtering runs. The track results demonstrated that adaptive filtering is a challenging problem for current systems. Indeed, when using the F1 utility measure to evaluate performance, the "baseline" system which retrieves no documents was the most effective system overall. Comparison with batch filtering results show that setting an appropriate threshhold for when to retrieve a document is a critical, and difficult, task in adaptive filtering.

6.4 The High Precision Track

Track reports– (Buckley, 1998, 1999a)

TREC-6 was the first running of the high precision track. The task in the track was to retrieve ten relevant documents for a topic within five minutes (wall clock time). Users could not collaborate on a single topic, nor could the system (or user) have previous knowledge of the topic. Otherwise, the user was free to use any available resources as long as the five minute time limit was observed. The task is an abstraction of a common retrieval problem: quickly find a few good documents to get a feel for the topic area.

Since the track guidelines put no limits on who the user could be, an implicit assumption of the track is that the runs were performed by system experts. As such, the track provides an upper-bound on the effectiveness obtainable by the systems. The 5-minute time limit was selected so that the intrinsic effectiveness of the system, the system efficiency, and the user interface would all be tested by the task.

The TREC-6 high precision track used the same 50 topics and document set as used in the TREC-6 ad hoc task. Five groups participated in the HP track, submitting a total of 13 runs. The mean over 50 topics of the precision after ten documents were retrieved ranged from a high 0.6020 to a low of 0.3360. The least effective runs were a set of completely automatic runs submitted to see how automatic runs would fare; the results confirm that user involvement is indeed beneficial. However, the best result was a run in which the user simply provided yes/no relevance judgments as input for a sophisticated (automatic) relevance feedback algorithm. This suggests that user involvement does not need to be extensive

The TREC-7 high precision track used the same 50 topics and document set as used in the TREC-7 ad hoc task. Four groups participated, submitting a total of seven runs. One finding of the track was that retrieving 15 good documents is a simple enough task for current retrieval systems that disagreements between the searcher and the assessor regarding what constitutes a relevant document bounds performance. However, new time-based evaluation measures introduced in the track offered a possible solution.

6.5 The Interactive Track

Track reports– (Over, 1997, 1998, 1999)

One of the first tracks to be started in TREC, the interactive track studies text retrieval systems in interaction with users and is interested in the process as well as the results. Effectively supporting the users of a retrieval system has become an increasingly important problem as more and more text is made electronically accessible, and larger numbers of end users (as opposed to a relatively small group of trained intermediaries) perform searches. Yet designing retrieval experiments that can be fairly evaluated and that produce interpretable results when humans are included in the loop is especially challenging since it is difficult to isolate the effects of the different factors that contribute to overall effectiveness.

Interactive experiments include a third factor, the searcher, to the topic and retrieval system factors inherent in all retrieval experiments. An ideal experimental design tests all combinations of all settings of all factors with repetitions, but with human subjects such a design is not feasible within a single site and certainly not across sites. For example, the same user cannot perform a search for a topic more than once because the experience gained during the first search biases the second search, but logistics prevents randomly assigning searchers from one site to perform searches on another site's experimental system. Finding a sufficient number of subjects is also difficult: experience indicates that reliably detecting significant system effects requires relatively many searches. Unfortunately, reducing the number of required searches by narrowing the focus of the investigation makes generalizing any conclusions difficult.

Based on the lessons learned from the TREC-4 track on how difficult it was to fairly compare results in interactive experiments, the track concentrated on experimental design in TREC-5. Unfortunately, the final design was not decided until late in the TREC cycle, and only two groups were able to participate. However the same design was used in TREC-6.

The goal of the TREC-6 interactive track was to compare systems across sites. To this end, the track developed and employed a new method for comparing interactive IR systems across different sites. The method involved comparing the particular retrieval system used at a site (an *experimental system*) to a common *control system* that was also run at each site. The direct comparison between the experimental and control systems was used to derive a measure of how much better the experimental system was than the control, independent of topic, searcher, and any other site-specific effects. Different experimental systems could then be indirectly compared across sites relative to the common control.

The experiment used six slightly modified ad hoc topics and the *Financial Times* newspaper data as the document collection. The searcher task involved six searches (three on control, three on the experimental system) to find and save documents which taken together contained as many answers as possible to the question stated or implied by the topic. Nine participants used this experimental framework to pursue their own research goals, and to contribute data to a cross-site comparison of systems. The evaluation measures used were recall and precision defined in terms of the set of all possible answers as determined by NIST assessors. Participants also reported extensive data on the characteristics of each searcher and of each searcher's interactions with both the control and experimental system.

As a first step in analyzing the cross-site data, the best model for each site's results in terms of which factors and interactions to include was determined. Then a cross-site analysis of variance (ANOVA) was performed, which indicated that there was a significant difference between some systems. However, a multiple comparisons test (Tukey's), run to determine which systems differed, found no significant pair-wise differences.

The effectiveness of using a control system to remove the site effect from cross-site comparisons was an assumption of the track design and so could not be tested by it. Additional experiments before and after TREC-6 did address the effectiveness of the control (i.e., the equivalence of the direct and indirect comparison of systems) but neither confirmed nor refuted its effectiveness (Lagergren & Over, 1998; Swan & Allan, 1998). As a practical matter, it is difficult to justify the cost of adding a control system to an experimental design in the absence of clear positive evidence for its effectiveness.

The TREC-7 track used a similar experimental framework, but without the requirement to use the single control system. The framework both defined a common task for participants to perform and prescribed an experimental matrix. The search task used the title and description sections plus a special "Instances" section of eight ad hoc topics; the documents searched were the *Financial Times* collection from Disk 4. The topics each described a need for information of a particular type such that multiple distinct examples or instances of that information were contained in the document collection. The searchers job was to save documents covering as many distinct answers to the question as possible in a 15-minute time limit. The NIST assessor for the topic made a comprehensive list of instances from the documents submitted by the track. The effectiveness of the search was evaluated by the fraction of total instances for that topic covered by the search (instance recall) and the fraction of the documents retrieved in the search that contained an instance (instance precision). Participants were also required to collect demographic and psychometric data from the searchers, and to report extensive data on each searcher's interactions with the search systems.

The experimental matrix defined how searchers and topics were to be divided among the experimental and control systems. (Participants were free to choose whatever systems they wanted to serve as experimental and control. That is, the track did not attempt to coordinate cross-site comparisons or test particular hypotheses.) The matrix was based on a latin square design, which provides the desired uncontaminated estimate of the difference between the systems. The minimum experiment defined by the design required eight searchers, with each searcher performing four searches with each of the two systems. The eight-searcher minimum was imposed since the results of the TREC-6 track suggested that with eight topics at least eight searchers are required to obtain statistically significant results.

Eight groups participated in the interactive track, performing a total of ten experiments. Since comparison of systems across sites was not supported by the experimental design, the results of the track need to understood in the context of the particular research goals of the individual research groups.

6.6 The Query Track

(Track report– (Buckley, 1999b))

The query track was a new track in TREC-7 whose goal was to create a large query collection. The variability in topic performance makes it impossible to reach meaningful conclusions regarding query-dependent processing strategies unless there is a very large query set—much larger than the sets of 50 topics used in the TREC collections. The query track was designed as a means for creating a large set of different queries for an existing TREC topic set, topics 1–50.

Participants in the track created different types of queries from the topic statements and/or relevance judgments. A query of a given type was created for each of the 50 topics, forming one query set. Five different query types were used:

Very short: two or three words extracted from the topic statement.

Sentence: an English sentence based on the topic statement and the relevant documents.

Manual feedback: an English sentence based on reading 5–10 relevant documents only (by someone who doesn't know the topic statement).

Manual structured query: a manually constructed query based on the topic statement and relevant documents. The use of operators supported by the participant's system was encouraged. The TIPSTER DN2 format was used to represent the query structure.

Automatic structured query: a query constructed automatically from the topic statement and relevance judgments. TIPSTER DN2 format used to represent the query structure.

Participants exchanged the query sets they created with all other participants in the track, and all participants ran all query sets their system could support. The document set used for the runs was the documents on Disk 2 plus the AP collection on Disk 3. The retrieval results were submitted to NIST where all runs were judged and evaluated.

Since the track design included all groups running all query sets, a number of direct comparisons were possible. First, participants could see how effective their system was using their own queries. Second, they could see how effective their search component was when using other queries, and finally, participants could evaluate how effective their query construction strategies were by seeing how other groups fared with their queries.

Unfortunately, only two groups participated in the query track, too few to make any meaningful comparisons. The track will run again in TREC-8, with the hope that heightened awareness of the problems the query track is addressing will generate participation.

6.7 The Confusion Track

(Track report– (Kantor & Voorhees, 1997))

A confusion (or data corruption) track was run in TREC-4 and TREC-5 to investigate the problems with using "corrupted" data such as would come from OCR or speech input. The TREC-4 track followed the ad hoc task, but using only the category B data. This data was randomly corrupted at NIST using character deletions, substitutions, and additions to create data with a 10% and 20% error rate (i.e., 10% or 20% of the characters were affected). Note that this process is neutral in that it does not model OCR or speech input. Four groups used the baseline and 10% corruption level; only two groups tried the 20% level. As was somewhat expected, the 10% error rate did not hurt performance in general and the track results were somewhat inconclusive.

In TREC-5, the test data was actual OCR output of scanned images of the 1994 *Federal Register.* Five groups participated in the experiment designed to explore the effect different levels of OCR error has on retrieval performance. This time a new task was tried: known-item searching. In this task the participants searched for particular previously identified documents in three versions of documents. The three versions of the documents were the original documents, the documents that resulted after the originals were subjected to an optical character recognition (OCR) process with a character error rate of approximately 5%, and the documents produced through OCR with a 20% error rate (caused by down-sampling the image before doing the OCR). The five groups tried very different methods, with the group from the Swiss Federal Institute of Technology (ETH) (Ballerini et al., 1997) performing the best, using a type of expansion of possible candidate words to improve the best match score.

It was decided to migrate the confusion track to the speech area in TREC-6, where it was called the Spoken Document Retrieval (SDR) track. The SDR track is a successor to the confusion track in that it represents a different form of "corrupted" documents. Instead of retrieving documents that are the result of OCR, systems retrieved documents that were the result of speech recognition systems.

6.8 The Spoken Document Retrieval (SDR) Track

Track reports– (Garofolo, Voorhees, Stanford, & Jones, 1998; Garofolo, Voorhees, Auzanne, Stanford, & Lund, 1999)

The SDR track fosters research on retrieval methodologies for spoken documents (i.e., recordings of speech). It was run in both TRECs 6 and 7, using different document sets and different tasks.

The TREC-6 document set was a set of transcripts from 50 hours of broadcast news originally collected by the Linguistic Data Consortium for DARPA Hub-4 speech recognition evaluations (Garofolo, Fiscus, & Fisher, 1997). Three versions of the transcripts were used: a "truth" transcript that was hand-produced; a transcript produced by an IBM baseline speech recognition system; and a transcript produced by the participant's own speech recognition system. Document boundaries were given in the hand-produced transcript, and the same boundaries were used in the other two versions. While recognizing fifty hours of news presented a serious challenge to the speech systems, the resulting document set was small by retrieval standards, consisting of only 1451 stories.

Like the earlier confusion tracks, the task in the TREC-6 SDR track was a *known-item search.* In a known-item search, the goal was to retrieve a single specific document, rather than a set of relevant documents. The search simulates a user seeking a particular, half-remembered document. NIST created 50 topics, each designed to describe precisely one document. Half of the topics were created to target speech conditions, and half to target retrieval conditions. Within each half, half were designed to be easy and half difficult. Difficult speech conditions included background noise, non-native speakers, low-bandwidth channels, and the like. Difficult retrieval conditions included the use of synonyms (e.g., cinema for movie theater) and rare senses of common words (e.g., looking for the document describing cigarette pants when many stories were about cigarette smoking).

Thirteen groups submitted SDR track runs. The results suggested that speech recognition and IR technologies are sufficiently advanced to do a credible job of retrieving specific documents. The better systems were able to retrieve the target document at rank 1 over 70% of the time using their own recognizer transcripts, compared to the best performance on the truth transcripts of 78.7%. Search performance was a bigger factor in the overall results than recognition accuracy, although the best results were obtained by groups that included both speech and IR experts.

The TREC-7 track implemented a full ranked retrieval task. The document collection consisted of transcripts of approximately 100 hours of broadcast news programs, representing about 3000 news stories. Participants worked with four different versions of the transcripts: the *reference* transcripts, which were hand-produced and assumed to be perfect; the first *baseline* transcripts, which were produced by a baseline speech recognition system running at about 35% word error rate; a second set of baseline transcripts, produced by the baseline recognizer running at about 50% word error rate; and the *recognizer* transcripts, which were produced by the participant's own recognizer system. Document boundaries were given in the hand-produced transcripts, and the same boundaries were used in the other versions.

NIST created a set of 23 topics, which were used to search each of the versions of the transcripts. The different versions of the transcripts allowed participants to observe the effect of recognizer errors on their retrieval strategy. The different recognizer runs provide a comparison of how different recognition strategies affect retrieval. To make this comparison as complete as possible, participants were encouraged to retrieve using other groups' recognizer transcripts as well. These runs are called *cross-recognizer* runs.

Eleven groups participated in the TREC-7 SDR track. The results of the track displayed a linear correlation between the error rate of the recognition and a decrease in retrieval effectiveness, a correlation that was not present in last year's track that used a known-item search task. Not surprisingly, the correlation is stronger when recognizer error rate is computed over content-based words (e.g., named entities) rather than *all* words.

6.9 The Very Large Corpus (VLC) Track

Track reports– (Hawking & Thistlewaite, 1998; Hawking, Craswell, & Thistlewaite, 1999)

The VLC track explores how well retrieval algorithms scale to larger document collections. In contrast to the ad hoc task that uses a 2 GB document collection, the first running of the VLC track in TREC-6 used a 20 GB collection, while the TREC-7 track used a 100 GB document collection.

The TREC-6 track's corpus consisted of 7.5 million texts for a total of 20.14 GB of data, including the five TREC CDs; USENET news postings; Canadian and Australian Hansards; HTML-formatted documents including university websites, and laws and judgments from the Australian Attorney General's Department; and the *Glasgow Herald* and *Financial Times* newspapers. The TREC-6 ad hoc topics were used.

Because of the difficulty of obtaining sufficient relevance judgments for recall-based measures, the main effectiveness measure used for VLC runs was precision after 20 documents were retrieved. Also reported were query response time; data structure (e.g., inverted index) building time; and a cost measure of number of queries processed per minute per hardware dollar. Participants were required to submit two runs: one run over the entire VLC corpus and a second run over a baseline collection that consisted of a random 10% sample of the full corpus. The focus of the evaluation was on the ratio of the measures between the baseline and full corpus runs.

Seven groups submitted VLC track runs. All of the participants were able to complete the VLC task with the hardware available to them (i.e., no special hardware purchases were made for the track). Indeed, the major conclusion of the track is that current systems are able to obtain good (high precision) retrieval effectiveness on a 20 GB collection with reasonable resources. For example, one of the best runs, from the University of Waterloo (Cormack et al., 1998), retrieved an average of 12.8 relevant documents in the top twenty processing at the rate of 2678 queries per hour using a cluster of four commodity PCs.

The TREC-7 collection consisted of World Wide Web data that was collected by the Internet Archive (`http://www.archive.org`). The track used the TREC-7 ad hoc topics, and a set of relevance judgments produced by assessors at the Australian National University. Because of the difficulty of getting sufficient relevance judgments to accurately measure recall, the main effectiveness measure used for VLC runs was precision after 20 documents were retrieved.

To more accurately measure the effect size has on the retrieval systems used by the participants, the track provided 3 collections: the original 100 GB collections plus 1% and 10% subsamples. Participants indexed each of the three collections and ran the entire topic set on each. They then reported timing figures for each phase as well as the top 20 retrieved. The main evaluation measures were precision after 20 documents retrieved (the effectiveness measure); query response time (elapsed time as seen by the user); data structure (e.g., inverted index) building time (elapsed time as seen by the user); plus a combination timing measure that factored in the expense of the hardware used.

Seven groups participated in the TREC-7 VLC track, with six groups processing the entire 100GB corpus. The track demonstrated that processing a 100GB corpus is well within the capabilities of today's retrieval systems. Of particular note was the Multitext group that achieved sub-second query processing time while maintaining good retrieval effectiveness using hardware that cost under US$100,000.

6.10 The Natural Language Processing (NLP) Track

(Track report- (Strzalkowski & Jones, 1997))

The NLP track was started in TREC-5 to explore whether the natural language processing techniques available today are mature enough to have an impact on IR, and specifically whether they can offer an advantage over more conventional methods. Four groups participated in the initial running of the natural language processing track.

The TREC-6 track used the 50 TREC-6 ad hoc topics and a reduced document set consisting of just the *Financial Times* newspaper data. The track had limited participation, with just two groups submitting NLP runs.

To date, specific NLP processing has not proved essential to obtaining effective retrieval in TREC. The most useful NLP techniques for text retrieval generally have been methods that recognize and normalize names and other multi-word terms. However, the TREC topics do not require processing at this level of detail. Other information seeking tasks such as fact extraction or story summarization may be a more appropriate test of current NLP technology.

6.11 The Database Merging Track

(Track report- (Voorhees, 1997))

The database merging track had the goal of investigating techniques for merging results from the various TREC subcollections (as opposed to treating the collections as a single entity). This type of investigation is important for real-world collections, and also to allow researchers to take advantage of possible variations in retrieval techniques for heterogeneous collections.

The track' was started in TREC-4, with 3 participating groups. running the ad hoc topics separately on each of the 10 subcollections, merging the results, and then submitting these, along with a baseline run treating the subcollections as a single collection. The 10 subcollections were defined corresponding to the various dates of the data, i.e., the three different years of the *Wall Street Journal*, the two different years of the *AP* newswire, the two sets of Ziff documents (one on each disk), and the three single subcollections (the *Federal Register*, the *San Jose Mercury News*, and the U.S. Patents).

If results are produced without use of collection information, then the merging process is trivial. Certainly this is one method of handling the problems of merging results from different databases. However this precludes using information about the collection to modify the various algorithms in the search engine, and, even more importantly, it does not deal with the issue about which collection to select. An implied question in this track was the hypothesis that one might want to bias searching towards certain collections.

There was a second running of the database merging track in TREC-5, again with only three groups participating. This time the data was split into many more (98) databases, to allow testing of database selection methods. Unfortunately this proved to be a high-overhead track and thus did not attract much participation despite a general interest in the problem. The track has not been run since TREC-5.

7 THE FUTURE

The final session of each TREC workshop is a planning session for future TRECs—especially to decide on the set of tracks for the next TREC. Two new tracks are planned for TREC-8, a question answering track and a Web track. The question answering track is designed to encourage research on methods for *information* retrieval as opposed to document retrieval. The goal in the track will be for systems to produce short text extracts that contain the answer for each of a set of 200 questions. The goal in the Web track will be to investigate whether links can be used to enhance retrieval. The track will use a 2GB subset of the data collected for the VLC track and a typical TREC ad hoc task. Also, participation in the query track is encouraged, since the benefits of that track increase with increased participation.

Acknowledgments

The authors gratefully acknowledge the continued support of the TREC conferences by the Intelligent Systems Office of the Defense Advanced Research Projects Agency. Thanks also go to the TREC program committee and the staff at NIST. The TREC tracks could not happen without the efforts of the track coordinators; our special thanks to them.

References

Allan, J., Ballesteros, L., Callan, J., Croft, B., & Lu, Z. (1996). Recent Experiments with INQUERY. In D. K. Harman (Ed.), (pp. 49–63). (NIST Special Publication 500-236.)

Allan, J., Callan, J., Croft, B., Bellesteros, L., Broglio, J., Xu, J., & Shu, H. (1997). INQUERY at TREC-5. In E. Voorhees & D. Harman (Eds.), (pp. 119–132). (NIST Special Publication 500-238.)

Allan, J., Callan, J., Croft, W. B., Ballesteros, L., Byrd, D., Swan, R., & Xu, J. (1998). INQUERY does battle with TREC-6. In E. Voorhees & D. Harman (Eds.), (pp. 169–206). (NIST Special Publication 500-240.)

Allan, J., Callan, J., Sanderson, M., Xu, J., & Wegmann, S. (1999). INQUERY and TREC-7. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Ballerini, J.-P., Büchel, M., Domenig, R., Knaus, D., Mateev, B., Mittendorf, E., Schäuble, P., Sheridan, P., & Wechsler, M. (1997). SPIDER Retrieval System at TREC-5. In E. Voorhees & D. Harman (Eds.), (pp. 217–228). (NIST Special Publication 500-238.)

Beaulieu, M., Gatford, M., Huang, X., Robertson, S., Walker, S., & Williams, P. (1997). Okapi at TREC-5. In E. Voorhees & D. Harman (Eds.), (pp. 143–166). (NIST Special Publication 500-238.)

Bodner, R., & Chignell, M. (1999). ClickIR: Text Retrieval using a Dynamic Hypertext Interface. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Boughanem, M., & Soulé-Dupuy, C. (1998). Mercure at Trec6. In E. Voorhees & D. Harman (Eds.), (pp. 321–328). (NIST Special Publication 500-240.)

Braschler, M., Krause, J., Peters, C., & Schäuble, P. (1999). Cross-Language Information Retrieval (CLIR) Track Overview. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Braschler, M., Wechsler, M., Mateev, B., Mittendorf, E., & Schäuble, P. (1999). SPIDER Retrieval System at TREC7. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Brown, E. (1995). Fast evaluation of structured queries for information retrieval. In *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 30–38).

Brown, E. W., & Chong, H. A. (1998). The GURU system in TREC-6. In E. Voorhees & D. Harman (Eds.), (pp. 535–540). (NIST Special Publication 500-240.)

Buckley, C. (1998). TREC-6 High-Precision Track. In E. Voorhees & D. Harman (Eds.), (p. 69-72). (NIST Special Publication 500-240.)

Buckley, C. (1999a). TREC-7 High-Precision Track. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Buckley, C. (1999b). TREC-7 Query Track. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Buckley, C., Mitra, M., Walz, J., & Cardie, C. (1998). Using Clustering and SuperConcepts within SMART: TREC-6. In E. Voorhees & D. Harman (Eds.), (pp. 107–124). (NIST Special Publication 500-240.)

Buckley, C., Mitra, M., Walz, J., & Cardie, C. (1999). SMART High Precision: TREC 7. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Buckley, C., Salton, G., Allan, J., & Singhal, A. (1995). Automatic Query Expansion Using SMART: TREC-3. In D. K. Harman (Ed.), (pp. 69–80). (NIST Special Publication 500-225.)

Buckley, C., Singhal, A., & Mitra, M. (1997). Using Query Zoning and Correlation Within SMART: TREC 5. In E. Voorhees & D. Harman (Eds.), (pp. 105–118). (NIST Special Publication 500-238.)

Clarke, C. L., & Cormack, G. V. (1997). Interactive Substring Retrieval (MultiText Experiments for TREC-5. In E. Voorhees & D. Harman (Eds.), (pp. 267–278). (NIST Special Publication 500-238.)

Cormack, G. V., Clarke, C. L., Palmer, C. R., & To, S. S. L. (1998). Passage-based refinement (MultiText experiments for TREC-6. In E. Voorhees & D. Harman (Eds.), (pp. 303–319). (NIST Special Publication 500-240.)

Evans, D., Huettner, A., Tong, X., Jansen, P., & Bennett, J. (1999). Effectiveness of Clustering in Ad Hoc Retrieval. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Fitzpatrick, L., & Dent, M. (1997). Automatic feedback using past queries: Social searching? In *Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 306–313).

Fuller, M., Kaszkiel, M., Ng, C. L., Vines, P., Wilkinson, R., & Zobel, J. (1998). MDS TREC6 report. In E. Voorhees & D. Harman (Eds.), (pp. 241–257). (NIST Special Publication 500-240.)

Garofolo, J., Fiscus, J., & Fisher, W. (1997). Design and preparation of the 1996 Hub-4 broadcast news benchmark test corpora. In *Proceedings of the DARPA speech recognition workshop* (pp. 15–21).

Garofolo, J., Voorhees, E., Auzanne, C., Stanford, V., & Lund, B. (1999). 1998 TREC-7 Spoken Document Retrieval Track Overview and Results. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Garofolo, J., Voorhees, E., Stanford, V., & Jones, K. S. (1998). 1997 TREC-6 Spoken Document Retrieval Track Overview and Results. In E. Voorhees & D. Harman (Eds.), (p. 83-92). (NIST Special Publication 500-240.)

Gey, F. C., Chen, A., He, J., Xu, L., & Meggs, J. (1997). Term importance, Boolean conjunct training, negative terms, and foreign language retrieval: probabilistic algorithms for TREC-5. In E. Voorhees & D. Harman (Eds.), (pp. 181–190). (NIST Special Publication 500-238.)

Harman, D. (1996). Overview of the fourth Text REtrieval Conference (TREC-4). In D. K. Harman (Ed.), (pp. 1–23). (NIST Special Publication 500-236.)

Harman, D. K. (Ed.). (1994, March). *Proceedings of the second text REtrieval conference (TREC-2).* (NIST Special Publication 500-215.)

Harman, D. K. (Ed.). (1996, October). *Proceedings of the fourth text REtrieval conference (TREC-4).* (NIST Special Publication 500-236.)

Hawking, D., Craswell, N., & Thistlewaite, P. (1999). Overview of TREC-7 Very Large Collection Track. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Hawking, D., & Thistlewaite, P. (1998). Overview of TREC-6 Very Large Collection Track. In E. Voorhees & D. Harman (Eds.), (p. 93-106). (NIST Special Publication 500-240.)

Hawking, D., Thistlewaite, P., & Bailey, P. (1997). ANU/ACSys TREC-5 Experiments. In E. Voorhees & D. Harman (Eds.), (pp. 359–376). (NIST Special Publication 500-238.)

Hawking, D., Thistlewaite, P., & Craswell, N. (1998). ANU/ACSys TREC-6 experiments. In E. Voorhees & D. Harman (Eds.), (pp. 275–290). (NIST Special Publication 500-240.)

Hiemstra, D., & Kraaij, W. (1999). Twenty-One at TREC-7: Ad-hoc and Cross-language Track. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Hull, D., Grefenstette, G., Schulze, B., Gaussier, E., Schütze, H., & Pedersen, J. (1997). Xerox TREC-5 Site Report: Routing, Filtering, NLP, and the Spanish Tracks. In E. Voorhees & D. Harman (Eds.), (pp. 167–180). (NIST Special Publication 500-238.)

Hull, D. A. (1998). The TREC-6 Filtering Track: Description and Analysis. In E. Voorhees & D. Harman (Eds.), (p. 45-68). (NIST Special Publication 500-240.)

Hull, D. A. (1999). The TREC-7 Filtering Track: Description and Analysis. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Kantor, P., & Voorhees, E. (1997). Report on the TREC-5 Confusion Track. In E. Voorhees & D. Harman (Eds.), (pp. 65–74). (NIST Special Publication 500-238.)

Kwok, K. (1996). A new method of weighting query terms. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 187–196).

Kwok, K., & Grunfeld, L. (1994). TREC-2 Document Retrieval Experiments using PIRCS. In D. K. Harman (Ed.), (pp. 233–242). (NIST Special Publication 500-215.)

Kwok, K., & Grunfeld, L. (1997). TREC-5 English and Chinese Retrieval Experiments using PIRCS. In E. Voorhees & D. Harman (Eds.), (pp. 133–142). (NIST Special Publication 500-238.)

Kwok, K., Grunfeld, L., & Xu, J. (1998). TREC-6 English and Chinese retrieval experiments using PIRCS. In E. Voorhees & D. Harman (Eds.), (pp. 207–214). (NIST Special Publication 500-240.)

Lagergren, E., & Over, P. (1998). Comparing interactive information retrieval systems across sites: The trec-6 interactive track matrix experiment. In *Proceedings of the 21th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 164–172).

Lewis, D. (1997). The TREC-5 Filtering Track. In E. Voorhees & D. Harman (Eds.), (pp. 75–96). (NIST Special Publication 500-238.)

Lu, A., Ayoub, M., & Dong, J. (1997). Ad Hoc Experiments using EUREKA. In E. Voorhees & D. Harman (Eds.), (pp. 229–240). (NIST Special Publication 500-238.)

Lu, A., Meier, E., Rao, A., Miller, D., & Pliske, D. (1998). Query processing in TREC6. In E. Voorhees & D. Harman (Eds.), (pp. 567–576). (NIST Special Publication 500-240.)

Mandala, R., Tokunaga, T., Tanaka, H., Okumura, A., & Satoh, K. (1999). Ad Hoc Retrieval Experiments Using WordNet and Automatically Constructed Thesauri. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Mateev, B., Munteanu, E., Sheridan, P., Wechsler, M., & Schäuble, P. (1998). ETH TREC-6: Routing, Chinese, cross-language, and spoken document retrieval. In E. Voorhees & D. Harman (Eds.), (pp. 623–635). (NIST Special Publication 500-240.)

Milic-Frayling, N., Evans, D., Tong, X., & Zhai, C. (1997). CLARIT Compound Queries and Constraint-Controlled Feedback in TREC-5. In E. Voorhees & D. Harman (Eds.), (pp. 315–334). (NIST Special Publication 500-238.)

Milic-Frayling, N., Zhai, C., Tong, X., Jansen, P., & Evans, D. A. (1998). Experiments in query optimization: The CLARIT system TREC-6 report. In E. Voorhees & D. Harman (Eds.), (pp. 415–454). (NIST Special Publication 500-240.)

Miller, D., Leek, T., & Schwartz, R. (1999). A hidden markov model information retrieval system. In *Proceedings of the 22th annual international ACM SIGIR conference on research and development in information retrieval* (p. TBD).

Nakajima, H., Takaki, T., Hirao, T., & Kitauchi, A. (1999). NTT DATA at TREC-7: system approach for ad hoc and filtering. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Namba, I., Igata, N., Horai, H., Nitta, K., & Matsui, K. (1999). Fujitsu Laboratories TREC7 Report. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Over, P. (1997). TREC-5 Interactive Track Report. In E. Voorhees & D. Harman (Eds.), (pp. 29–56). (NIST Special Publication 500-238.)

Over, P. (1998). TREC-6 Interactive Track Report. In E. Voorhees & D. Harman (Eds.), (pp. 73–81). (NIST Special Publication 500-240.)

Over, P. (1999). TREC-7 Interactive Track Report. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Pedersen, J. O., Silverstein, C., & Vogt, C. C. (1998). Verity at TREC-6: Out-of-the-box and beyond. In E. Voorhees & D. Harman (Eds.), (pp. 259–273). (NIST Special Publication 500-240.)

Rao, A., Humphrey, T., Parhizgar, A., Wilson, C., & Pliske, D. (1999). Experiments in Query Processing at LEXIS-NEXIS for TREC-7. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Robertson, S., Walker, S., Hancock-Beaulieu, M., & Gatford, M. (1994). Okapi and TREC-2. In D. K. Harman (Ed.), (pp. 21–34). (NIST Special Publication 500-215.)

Salton, G., & McGill, M. (Eds.). (1983). *Introduction to modern information retrieval.* McGraw-Hill Book Co., New York, NY.

Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology, 29*, 3–48.

Schäuble, P., & Sheridan, P. (1998). Cross-Language Information Retrieval (CLIR) Track Overview. In E. Voorhees & D. Harman (Eds.), (pp. 31–43). (NIST Special Publication 500-240.)

Singhal, A. (1998). AT&T at TREC-6. In E. Voorhees & D. Harman (Eds.), (pp. 215–225). (NIST Special Publication 500-240.)

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 21–29).

Singhal, A., Choi, J., Hindle, D., Lewis, D., & Pereira, F. (1999). AT&T at TREC-7. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Smeaton, A., & Wilkinson, R. (1997). Spanish and Chinese Document Retrieval in TREC-5. In E. Voorhees & D. Harman (Eds.), (pp. 57–64). (NIST Special Publication 500-238.)

Sparck Jones, K. (in press). Further Reflections on TREC. *Information Processing and Management.*

Sparck Jones, K., & van Rijsbergen, C. (1975). *Report on the need for and provision of an "ideal" information retrieval test collection.* British Library Research and Development Report 5266. Computer Laboratory, University of Cambridge.

Strzalkowski, T., & Jones, K. S. (1997). NLP Track at TREC-5. In E. Voorhees & D. Harman (Eds.), (pp. 97–102). (NIST Special Publication 500-238.)

Strzalkowski, T., Lin, F., & Perez-Carballo, J. (1998). Natural Language Information Retrieval: TREC-6 Report. In E. Voorhees & D. Harman (Eds.), (pp. 347–366). (NIST Special Publication 500-240.)

Strzalkowski, T., Lin, F., Wang, J., Guthrie, L., Leistensnider, J., Wilding, J., Karlgren, J., Straszheim, T., & Perez-Carballo, J. (1997). Natural Language Information Retrieval: TREC-5 Report. In E. Voorhees & D. Harman (Eds.), (pp. 291–314). (NIST Special Publication 500-238.)

Strzalkowski, T., Stein, G., Wise, G. B., Perez-Carballo, J., Tapananinen, P., Jarvinen, T., Voutilainen, A., & Karlgren, J. (1999). Natural Language Information Retrieval: TREC-7 Report. In E. Voorhees & D. Harman (Eds.), (p. TBD). (NIST Special Publication 500-242.)

Swan, R., & Allan, J. (1998). Aspect windows, 3-d visualizations, and indirect comparisons of information retrieval systems. In *Proceedings of the 21th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 173–181).

Voorhees, E. (1997). The TREC-5 Database Merging Track. In E. Voorhees & D. Harman (Eds.), (pp. 103–104). (NIST Special Publication 500-238.)

Voorhees, E. (in press). Special issue: The sixth Text REtrieval Conference (TREC-6). *Information Processing and Management.*

Voorhees, E., & Harman, D. (Eds.). (1997, November). *Proceedings of the fifth Text REtrieval Conference (TREC-5).* (NIST Special Publication 500-238.)

Voorhees, E., & Harman, D. (Eds.). (1998, August). *Proceedings of the sixth Text REtrieval Conference (TREC-6).* (NIST Special Publication 500-240.)

Voorhees, E., & Harman, D. (Eds.). (1999, April). *Proceedings of the seventh Text REtrieval Conference (TREC-7).* (NIST Special Publication 500-242.)

Voorhees, E., & Harman, D. (in press). Overview of the Sixth Text REtrieval Conference (TREC-6). *Information Processing and Management.*

Voorhees, E. M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21th annual international ACM SIGIR conference on research and development in information retrieval* (p. 315-323).

Walker, S., Robertson, S., Boughanem, M., Jones, G., & Sparck Jones, K. (1998). Okapi at TREC-6: Automatic ad hoc, VLC, routing, filtering and QSDR. In E. Voorhees & D. Harman (Eds.), (pp. 125–136). (NIST Special Publication 500-240.)

Wilkinson, R. (1998). Chinese Document Retrieval at TREC-6. In E. Voorhees & D. Harman (Eds.), (pp. 25–30). (NIST Special Publication 500-240.)

Xu, J., & Croft, W. (1996). Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 4–11).

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments. In *Proceedings of the 21th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 307–314).